DDN® STORAGE

+

intel®

WHITEPAPER

DDN, INTEL COLLABORATION RAMPS UP NEXT GENERATION SEQUENCING

KETAN PARANJAPE, *General Manager Life Sciences, Intel Corporation*
DR. GEORGE VACEK, *Global Business Director Life Sciences, DDN*
LAURA SHEPARD, *Sr. Director Vertical Markets, DDN*

When it comes to generating increasingly larger data sets and stretching the limits of high performance computing (HPC), the field of genomics and next generation sequencing (NGS) is in the forefront.

The major impetus for this data explosion began in 1990 when the U.S. kicked off the Human Genome Project, an ambitious project designed to sequence the three billion base pairs that constitute the complete set of DNA in the human body. Eleven years and $3 billion later the deed was done. This breakthrough was followed by a massive upsurge in genomics research and development that included rapid advances in sequencing using the power of HPC. Today an individual's genome can be sequenced overnight for less than $1,000.

DEALING WITH BIG DATA

But these gains come with their share of challenges. Big data is one of them. Technological advances allow today's sequencers to generate 4x the data in 50% less time than from only a couple of years ago – on the order of one TB per device per day. In microscopy 4D, molecular imaging products are generating 2TB per device per day. Working with a fragmented software ecosystem that relies heavily on open source contributions, IT must ingest, store, manage, move and share petabytes of research and clinical data. Reliable snapshots of pipelines must be generated and archived to tiered storage. And there is a growing need for secure data sharing to foster collaborations that span departments, organizations, and countries.

Rapid iterations of algorithms place further stress on IT, requiring the utmost in flexibility and agility. But most applications do not fully leverage the available infrastructure. Average CPU utilization is low – most cores are not being used. The same is true for I/O bandwidth to memory. In addition, the average memory footprint is small – applications are not using the memory available in newer HPC systems powered by processors and coprocessors such as the Intel® Xeon® product family and Intel® Xeon Phi™ coprocessor.

Due to the constantly growing genomics datasets and the need to analyze this big data in close to real time, storage is a major consideration, not only from the perspective of capacity, but also from the perspective of IOPs and bandwidth to disk. In fact, the major roadblocks to key life sciences applications – such as NGS-based personalized medicine – have more to do with the available computing infrastructure than the complexity of the genetic code. Compute, fabric, file systems and storage must be optimally balanced and scale equally to create an IT infrastructure that can keep pace with the revolution in genomics.

TOWARD A SOLUTION

This is why Intel and DataDirect Networks (DDN) have teamed up to create a High Performance infrastructure that fully supports the creation and management of genomic data now and in the future.

For its part, Intel is working with life sciences industry experts worldwide targeting both Intel Xeon processors and Intel Xeon Phi coprocessors to deliver optimized performance from top genomics applications. Intel scientists and engineers help implement fine- and coarse-grained optimization at the node and cluster level, and work with code authors to release the optimizations and disseminate best practices.

FOCUS ON STORAGE

In order to analyze and realize value from the huge dataflows generated by these applications, organizations require a storage infrastructure that is highly scalable, and can support I/O thirsty applications for high-throughput data processing.

About half of today's enterprise storage systems are based on NAS. Unfortunately, NAS does not perform well at scale. This spurred the development of Lustre*, the very popular open source parallel file system that is highly scalable and in use wherever HPC

and huge amounts of storage capacity are required.

Intel has released a new generation of Lustre software –. Intel Enterprise Edition for Lustre software, a hardened, commercial grade version of Lustre optimized to handle HPC storage and throughput challenges.

Intel EE for Lustre software enables fully parallel I/O across clients, servers and storage devices. Massive data flows are able to use a high percentage of the underlying storage and networking fabric for low latency, high-throughput storage performance. This allows organizations to run larger and more complex simulations faster and more easily.

A native Lustre client optimized for Intel Xeon Phi coprocessor delivers data 10x faster than NFS – a must for handling the huge data flows associated with life sciences research. Also, Intel EE for Lustre software scales up to tens of thousands of clients and petabytes of data. The largest current Lustre implementation has 25,000 clients attached to it at the same time.

Today data can be accessed through Lustre at sustained speeds of 2TB/s, with production customers – like Oakridge National Lab - using Lustre on DDN at over 1 TB/s. By the end of the decade that number should reach 10TB/second.

DDN AND INTEL

DDN configures, sells and installs more Lustre* than any other storage vendor, and is the largest reseller of Intel EE for Lustre software. DDN is the leading supplier of high performance data storage systems to the genomics community, powering more than one third of the top sequencing centers.

DDN's EXAScaler® appliances, based on its Storage Fusion Architecture® (SFA), combine DDN's industry leading performance and density with Intel EE for Lustre software. This approach creates a turnkey solution that delivers high performance and is easy to size, install, manage and grow.

DDN's SFA® product family uses the most advanced Intel processor technology, bus and memory with an optimized RAID engine and sophisticated data management algorithms. It combines SATA, SAS, and solid-state disks (SSDs) for an environment that can be tailored to balance throughput, capacity, scale up, or scale out as needed.

SFA consists of 84 drives in each 4U enclosure that can scale up to 7PB per rack. By leveraging the IOPS capabilities of the SSDs, the flagship SFA platform – the SFA14K™ – delivers up to 60 GB/sec throughput and up to 6 million IOPS.

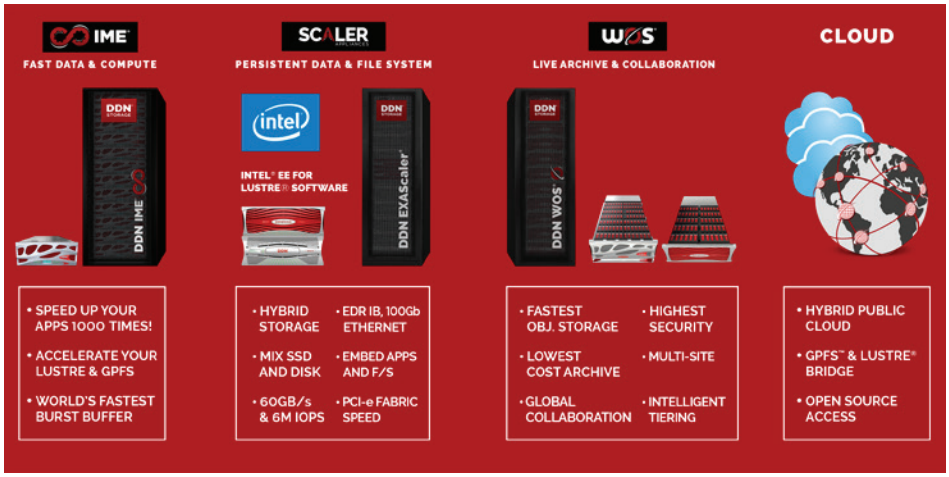
EXAScaler customers have reported up to 16x faster processing of genomic data. Other DDN genomics customers have documented massive increases to their big data genomics pipelines and far simpler to scale as data sets increase from TB to PB in size. For example, at Wellcome Trust Sanger Institute, 30 Illumina sequencers are generating more than 36TB a week, a torrent of big data that is easily handled by their EXAScaler appliance with its 22.5PB of usable storage and a sustained data rate of 20GB/s.

*Some names and brands may be claimed as the property of others.

DELIVERING END-TO-END DATA LIFECYCLE MANAGEMENT SOLUTIONS

Fast ingest of and access to ever larger data sets is critical to sequencing workflows—but so are simple, cost-effective, long term data retention and collaboration.

DDN Scaler Parallel File System Appliances with Intel EE for Lustre share a unified namespace with other storage tiers for a single application, administrator and end user view of data regardless of whether it resides on high performance storage, active archive, tape or cloud.



WOS AND THE PRIVATE CLOUD

DDN WOS® (web object scaler) is a scale out object-based storage system that allows organizations to quickly and easily build and deploy private storage clouds across geographically distributed sites. It allows genomics researchers to expand their IT infrastructure for collaboration and publication of sequence data and analysis.

WOS has proven to be an outstanding platform for content distribution and collaboration clouds as well as for active archives, high availability and disaster recovery. It delivers control over the performance and cost of big data storage and access, regardless of the data's physical location.

In addition to providing true object storage, WOS also features a federated global namespace that can be built out to exabyte dimensions. Other features include flexible data protection, a self-healing architecture and a latency-aware access manager that minimizes latency for end users. WOS also includes metadata search capabilities that avoid scalability bottlenecks and allow users to query metadata across petabytes of objects.

WOS includes the highest performance results for storage performance in terms of throughput, IOPS and latency. A single WOS cluster can provide up to 3 million IOPS and 300 GB/s throughput. Multiple clusters can be combined for even higher performance.

JAPANESE BIOBANK LEADS THE WAY WITH LARGE POPULATION STUDIES OF MULTI-GENERATION GENOMIC AND PHENOTYPIC DATA

The Tohoku Medical Megabank Organization (ToMMo) at Tohoku University is making huge strides in genomics and human health. ToMMo was founded to establish an advanced medical system supporting health and welfare in the Tohoku area, as part of the reconstruction after the Great East Japan Earthquake.

The research at ToMMo's biobank combines genomic and medical data of hundreds of thousands of patients from the region. One of the biggest challenges in epidemiological research is study size – studies need to recruit very large cohorts to ensure that results are statistically significant – and the biobank at ToMMo has been very successful in recruiting participants. Because of the multi-generational families in the area, the bank includes cohort data that spans three generations

for studies into the relationship of genetic and environmental factors on disease susceptibility and therapeutic effectiveness. The research informs policy decisions and evidence-based practice by identifying risk factors for disease and targets for preventive healthcare.

The importance of these studies and insight cannot be underestimated. ToMMo has already completed high-accuracy sequencing of 1,000 individuals. Managing all of this structured and unstructured data is critical. And, in this type of environment, a number of things became imperative for ToMMo scientists and researchers: their data had to be highly available; they wanted to be able to accelerate the performance of their systems in order to maximize processing time for sequencing data; and, they needed to be able to scale the amount of storage to support anticipated growth in high throughput sequencing and analysis.

For ToMMo, this meant deploying one of the largest Lustre systems in the world dedicated to genome research, with 16.6 PB of DDN storage that is expandable to more than 50 PB as they continue to grow the biobank over the next few years. ToMMo now has the infrastructure and tools needed to support their medical genomics with unparalleled performance and scale, as well as being able to reap the integration benefits of sourcing both high performance file system and cloud storage from one single vendor.

GENOME RESEARCH INSTITUTE REDUCES STORAGE COMPLEXITY AND ACCELERATES RESEARCH EFFORTS

The Wellcome Trust Sanger Institute, a charitably funded genomic research center located in the United Kingdom, is a world leader in studying the impact of genetics and genomics on global health. Since its inception in 1993, Sanger Institute has developed new understanding of genomes and their role in biology while delivering some of the most important advances in genomic research.

When it came to handling its Big Data requirements, the Institute faced a number of challenges.

Obtaining a robust IT infrastructure with large-scale, high-throughput performance is essential to supporting Sanger Institute's diverse research community, encompassing over 2,000 scientists worldwide. Major sequencing technology advancements, including powerful new machines, created a surge in data volume and computational analysis.

The problem was complicated by unpredictable data growth as the amount of data and computational analysis varied by workload and research project. Delivering on the extremely high service levels that were required to store, manage, access and archive massive amounts of research data.

Sanger's solution included DDN SFA high-performance storage engine and EXAScaler Lustre file system appliance to deliver unprecedented levels of throughput and scalability to support tens of thousands of data sequences requiring up to 10,000 CPU hours of computational analysis.

As a result, DDN enabled Sanger Institute further its exploration of ground-breaking scientific and medical research. Sanger Institute is now well positioned to keep pace with advancements in sequencing technology with storage that can scale seamlessly without replacement or forklift upgrades.

DDN enables Sanger Institute to achieve its business goal of ensuring open data sharing by making it easy for the worldwide scientific community to collaborate and access the latest data and analytics.

INTEL® AND DDN: A WINNING COMBINATION

The combination of Intel processors and coprocessors, Intel Enterprise Edition for Lustre software and DDN storage appliances provides a powerful platform for the ingest, analysis, search, collaboration and archiving of massive amounts of genomic data.

The joint efforts of the two companies take full advantage of their deep experience in the life sciences and key related technologies. Together Intel and DDN are accelerating genomics imaging and modeling workflows to drive more actionable results in less time for more researchers than any other solution.

ABOUT INTEL

Intel® is helping drive the life sciences evolution through a comprehensive approach that includes working with key players in the industry. Together, we are pursuing an end-to-end solution to the computing challenges faced in personalized care and the development of new plant and drug compounds. We draw on expertise spanning the entire compute continuum, from local high performance computing (HPC) clusters, big data analytics, cloud, and commercial and open source software initiatives, to mobile devices, smartphones, and sensors.

ABOUT DDN®

DataDirect Networks (DDN) is the world's leading big data storage supplier to data-intensive, global organizations. For more than 15 years, DDN has designed, developed, deployed and optimized systems, software and solutions that enable enterprises, service providers, universities and government agencies to generate more value and to accelerate time to insight from their data and information, on premise and in the cloud. Organizations leverage the power of DDN technology and the deep technical expertise of its team to capture, store, process, analyze, collaborate and distribute data, information and content at largest scale in the most efficient, reliable and cost effective manner. DDN customers include many of the world's leading financial services firms and banks, healthcare and life science organizations, manufacturing and energy companies, government and research facilities, and web and cloud service providers. For more information, visit our website www.ddn.com or call 1-800-837-2298.

DDN® STORAGE

BOTTOM LINE RESULTS. ACCELERATED.

SALES@DDN.COM | +1.800.837.2298

v3 (1/16)