Case Study
Intel® Xeon® Processor E5 Series
Intel® Solid-State Drives
High Performance Computing

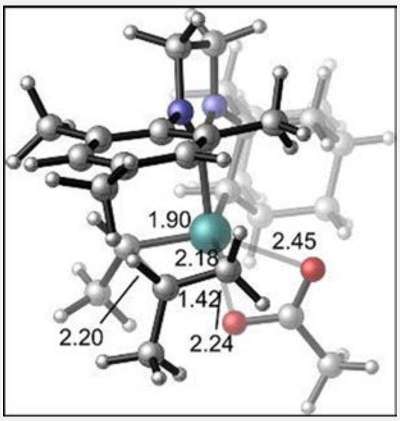
Flash Technology in High-Performance Computing Accelerates Scientific Discovery

Intel® Xeon® processor E5 series and Intel® Solid-State Drives help to improve the research productivity of the computational chemist and biologist.

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

"With outstanding performance, Intel® SSDs help SDSC's researchers perform ground-breaking scientific research. Flash provides an order of magnitude improvement in I/O over traditional HPC disk architectures."

Michael Norman
Executive Director
San Diego
Supercomputer Center



Z-selective olefin metathesis transition state (source K. Houk, UCLA)

INTRODUCTION

SDSC is a leader in data-intensive computing, providing resources, services and expertise to the national research community including industry and academia. The mission of SDSC is to extend the reach of scientific accomplishments by providing tools such as high-performance hardware technologies, integrative software technologies, and deep interdisciplinary expertise to these communities. In 2012 SDSC deployed *Gordon,* the first HPC cluster to use flash on a massive scale. *Gordon* was built by Appro International, which has HPC deployments in universities, government research labs, and companies around the world.

THE I/O CHALLENGE

Today, many scientific and data intensive applications are characterized by data access patterns that are random in nature, or where there is a high frequency of very small transactions. Conventional HPC architectures, which provide large scale parallel file systems (e.g., Lustre) built on spinning disk technology, are not well-suited for this type of I/O. The associated file system and mechanical latencies are simply too high to support the I/O transactions as applications move data between the parallel file system and the CPU.

SOLUTION

Gordon inserts a new layer in the memory hierarchy by deploying 300 TB of Intel® SSD 710 Series flash between the parallel file system and the compute nodes. Sixty four Intel Westmere I/O nodes, each with 16 Intel 710 Series SSDs provide users with a resource that can be used as fast scratch space, or as persistent storage for database and analytics application.

Use Case #1: Eliminating I/O Bottlenecks for Computational Chemistry

Kendall Houk and Pen Liu UCLA, and Bob Grubbs, Caltech, are working to find new catalysts for the process of olefin metathesis, which is important in the development of industrial and consumer products. For this work, they are using Gaussian, the de facto standard application for performing computational chemistry calculations. Gaussian creates large temporary scratch files to store and retrieve interim results, and can consume a large amount of memory. These requirements push the limits of traditional HPC systems, and up until now, have severely limited the scale and complexity of the calculations that are possible on these systems.

Gordon uses Intel® SSDs to provide a new level in the memory hierarchy. This is enabling new classes of applications and scales of computation not previously possible on HPC architectures.

Spotlight on Gordon

- 1,024 2S compute nodes with Xeon E5 2.6 GHz, 8-core processors, 64 GB/node
- 300 TB of Intel 710 SSDs
- Dual rail 3D torus interconnect
- Large memory nodes using ScaleMP's vSMP Foundation

But with Gordon, having 64 GB of memory, and 300 GB of flash on each Xeon E5 compute note, these calculations are now routine. Houk notes, "Using flash disk and large amounts of shared-memory has dramatically increased the performance of the calculations and enabled more accurate computational methods such as CCSD(T) on larger molecules". He goes on to state, "This computational method is generally referred to as the 'gold standard' for accuracy by a single calculation. This example requires more than 100GB scratch space, and this is only practical with the SSD local scratch on *Gordon*."

SDSC is seeing a huge increase in allocation requests tied to Gaussian and Gordon's flash capability and for good reason -- users are finding that they are now able to perform calculation at scales, and in number previously not possible.

Use Case #2: Enabling HPC Database Applications for Life Science Research

The Protein Data Bank (PDB) archive is the single worldwide repository of information about the 3D structures of large biological molecules. These are the molecules that are found in all living organisms. The 3-dimensional arrangement of each of these proteins provides the complexity and richness of life as we know it. Comparing proteins to one another is critical to many life science domains, including drug discovery. The PDB routinely carries out this comparison, but it is a computationally expensive task which can take over several weeks to complete on traditional HPC architectures. The bottleneck in this calculation is the underlying I/O performance of the database server which is built using standard spinning disks (HDD). Simply put, the latency of the spinning disk cannot keep up with the high transaction rates generated by the computation.

Phil Bourne, Co-PI of the RCSB PDB notes, "We first developed algorithms to do this in 1998, but lacked the computer power to perform all the needed pairwise comparisons."

Andreas Prlic, of the PDB, working with Pietro Cicotti, SDSC, have ported these algorithms to Gordon. The PDB MySQL database is installed on a *Gordon* flash-based I/O node, and subsequently used to dispatch compute takes to the Gordon compute nodes. With latencies that are an order of magnitude lower than HDD, the calculation proceeds much faster. Initial comparisons have shown a 4x improvement in speed over a traditional hard disk-based implementation. Calculations that were taking 67 days have been reduced to 18.

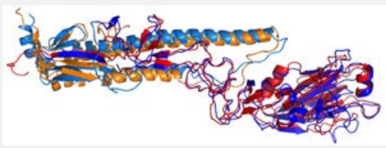


Figure 1A protein structure comparison of two influenza hemaggutinin molecules

Bourne notes, "For the first time we are in a position to accurately compare to each other all the known 3-dimensional shapes of these proteins to quantitatively characterize their similarities and differences. With *Gordon* we have been able to get a significant improvement in performance and are currently undertaking this comparison and expect to learn much from the analysis."

Gordon as a platform for testing new flash technology

Most recently, SDSC tested the new Intel® SSD DC S3700 Series, the follow-on to the 710's in Gordon. The DC S3700 provides twice the I/O performance of the 710, with 4KB random read performance up to 78,000 IOPS and tighter latency of <60uS, providing fast and consistent performance. The result is better than a 2x reduction in the I/O time for many HPC computations. The end-to-end AES data and power loss protection of the DC S3700 ensures that the data written is accurate—a capability that allows SDSC to deliver high reliability HPC infrastructure.

Find the Intel Solid-State Drive solution that is right for your business. Contact your Intel representative or visit www.intel.com/go/ssd for product information.

To learn more about other Intel business solutions, visit Business Success Stories for IT Managers at www.intel.com/itcasestudies

1111/JAY/PMG/XX/PDF 326293-001US