Solutions Brief

Scientific Workflow Solution for Computational Chemistry

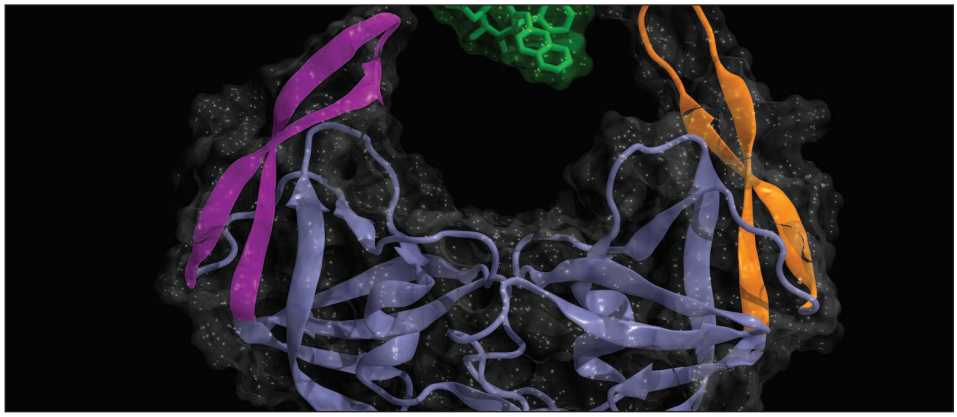


Image courtesy of Stony Brook University.

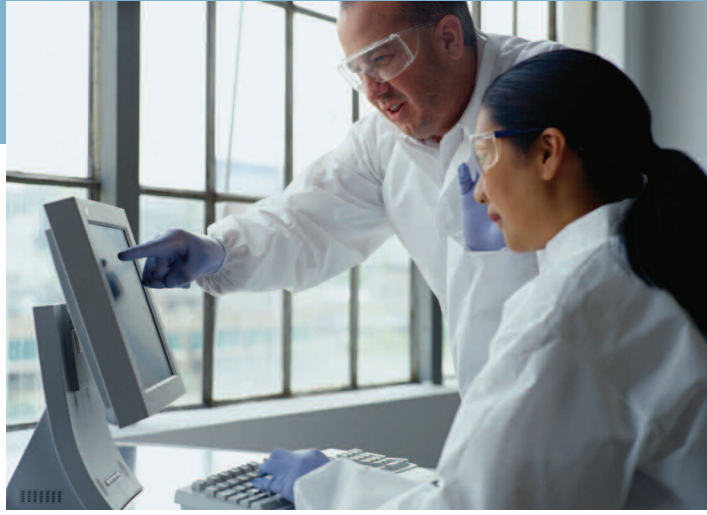
A Multi-System Platform for Modeling Biological and Materials Behavior

Simulating—and Solving—Critical Problems

Understanding the fundamental nature of molecules is crucial for scientific breakthroughs in numerous disciplines. Today's computational chemists rely on advanced algorithms and high-performance computers to develop and analyze models that simulate some of the most important problems faced in our world today. Whether predicting the interaction of ligands with proteins, studying reactions on surfaces, or developing new nanomaterials, scientists use computational chemistry simulations to discover fundamental relationships between the structure, properties and behavior of chemical compounds.

Computational chemistry environments in bioscience and materials science share a number of significant challenges, the solution of which allows scientists to focus on their science instead of computer science, makes individuals and groups more productive, and increases ROI.

- **Large and complex problems:** Scientists are simulating larger and larger proteins and materials at the quantum and molecular level. For example, scientists at NCSA recently using molecular dynamics to simulate an entire tobacco virus. These simulations are intractable on small systems and can take weeks or months to run on even the largest systems.
- **High throughput:** Pharmaceutical researchers routinely simulate millions of potential ligand-protein dockings so that time-consuming laboratory testing can focus on the most likely candidate drugs. Turning an overnight analysis into one completed interactively could increase the insight available to scientists and reduce development cycle-time and costs.
- **Exploding data sizes:** Scientists regularly use input databases and generate results that can exceed multiple terabytes in size. For instance, regularly used proteomics data bases are 8 terabytes in size, and are rapidly expanding. Working with full data sets instead of



partial problems would improve the productivity of individual researchers.

- **Complex user environments:** Computational chemists routinely utilize multiple systems and data storage environments, and they spend an increasing amount of time accessing data and managing where jobs run. A workflow-oriented approach that unifies system and data access would enable researchers to focus on chemistry instead of the intricacies of computer system access.

SGI's Scientific Workflow Solution for Computational Chemistry helps scientific enterprises overcome these challenges by integrating multiple computing and storage architectures into a flexible solution that can be tailored to optimally meet the current needs of any organization and evolve to match future requirements.

Multi-Application and Multi-Workflow Requirements

Multiple distinct workflows are often present within a single bioscience or materials science environment. Each of these workflows has its own computing requirements that change as individual projects progress. Some workflows are large-job intensive, others work with large databases, and others require high-throughput to run thousands or millions of small to medium size jobs.

The divergence among computing requirements found in a single organization means that the standardization on a single computing platform sub-optimally supports overall research goals, and in extreme cases eliminates certain types of critical calculations from the computational chemists' repertoire.

There are a myriad of computational chemistry application requirements present in today's high-performance, high-throughput workflows. Several key requirements include.

- **Scientific Databases –** with genomic, proteomic and structure databases reaching the multi-terabyte range, efficiently finding the right candidate compounds to work with in the rest of your workflow is of critical importance. SGI® Altix® systems have demonstrated breakthrough performance for medium and large traditional databases such as Oracle® 10g™ and IBM® DB2®, in-memory databases such as Oracle TimesTen, and the myriad proprietary databases often found in bioscience and materials science environments
- **Shape Comparison and Docking –** applications like ROCS, GLIDE and FLEXX are key elements of throughput-oriented workflows that determine which candidate compounds are further analyzed with quantum mechanics applications. Thousands of candidates must be tested with multiple conformations that result in hundreds of millions of analyses. SGI Altix XE cluster systems combine an integrated system management environment with multiple dual-core Intel Xeon® 5100 sequence processors that are especially well suited to running docking applications.
- **Molecular Mechanics and Molecular Dynamics (MM/MD) –** applications like Amber, NAMD and Gromacs are used to simulate the behavior of ever larger and more complex structures. These applications require both single CPU performance and extreme scalability. Depending on the specific problem being solved, these applications run well on SGI Altix systems, SGI Altix XE cluster systems, or both.
- **Quantum Mechanics (QM) –** semi-empirical QM applications like AMPAC utilize the high CPU performance available on both SGI Altix and SGI Altix XE systems to analyze thousands of compounds, and ab initio and plane-wave QM applications like Gaussian, CASTEP, DMol^3, Jaguar and VASP are optimized to achieve industry-leading CPU performance and scalability on SGI Altix systems. SGI systems are uniquely positioned to combine the benefits of quantum mechanics and molecular mechanics (QMMM) to allow users to increase the theory level and accuracy for critical regions of interest.

SGI's Scientific Workflow Solution for Computational Chemistry supports the optimal solution of each of these individual problems as well as an organization's diverse computational chemistry requirements.

A Multi-Workflow Architecture

The SGI Scientific Workflow Solution for Computational Chemistry is designed to support the different computing and data access requirements found in today's leading bioscience and materials science environments. This solution combines the high-performance capability of SGI Altix systems with the high-throughput capabilities of SGI Altix XE cluster systems and SGI's family of storage and data management platforms.

Every computational chemistry workflow is unique, and different organizations combine them in different ways. The end results are requirements that vary tremendously from site to site and change over time. The SGI Scientific Workflow Solution for Computational Chemistry is a modular solution that can

be configured to match any organization's needs. It also helps maximize the value of your investment by enabling you to add more high-performance computing, high-throughput computing, large database support, and data management capabilities without replacing your existing environment. Following are just a few representative examples:

- A workflow with a large amount of ab initio quantum mechanics and a small amount of small-molecule molecular dynamics will be more heavily weighted with the high-performance capabilities found in SGI Altix systems.
- A workflow with a large amount of docking and a small amount of large-molecule molecular dynamics will be more heavily weighted with the high-throughput capabilities found in SGI Altix XE cluster systems.
- A workflow with a large amount of large-memory database work would be primarily focused on SGI Altix systems and SGI® InfiniteStorage high-performance storage and archiving capability.

The SGI Scientific Workflow Solution for Computational Chemistry was designed to allow scientists to focus on their science. The solution gives them one point of access to both high-performance and high-throughput computing, and allows all systems – including desktop systems – to share one set of high-performance data storage.

Common system-level characteristics make it possible to move from one platform to another, and to manage projects that span multiple systems and technologies:

- A common Linux operating system can be provided across all platforms so that users need learn only one system.
- A single job-scheduling and workload management system can be used to manage workflow across all systems.
- SGI InfiniteStorage storage solutions simplify data sharing, access, movement across systems, and information lifecycle management within an organization.
- A centralized systems management interface and operational dashboard for updating, provisioning, configuring, and monitoring the solution.

SGI® Altix®: A High-Performance Shared-Memory Platform

SGI® Altix® servers provide industry leading compute performance and scalability for computationally intensive quantum mechanics applications, large-molecule molecular dynamics simulations and scientific databases. They also deliver the maximum flexibility which allows organizations to dynamically switch between running a broad mix of computational chemistry applications and extremely large breakthrough calculations.

Built using SGI's advanced NUMAflex™ technology, these mid-range and high-end systems utilize the Dual-Core Intel® Itanium® 2 9000 sequence processors and are able to scale from four CPU cores to 1,024 CPU cores in a single, shared memory environment. This design minimizes system overhead and makes it easy for applications to run faster in a scalable environment.

The end result is that applications like large scientific databases, quantum mechanics and large model molecular dynamics scale to new levels of performance, enabling faster insights and a deeper understanding of chemical processes. And, with mid-range systems scaling up to 76 CPU cores and over 400GBytes of memory, smaller and smaller organizations can solve problems that, until recently, could not be considered anywhere.

SGI Altix XE: A High-Throughput, Cluster Platform

SGI Altix XE cluster systems provide industry leading high-throughput capability for semi-empirical quantum mechanics, small molecule molecular dynamics, crystallography, docking, and a wide range of additional computational chemistry and bioscience applications. Many of these applications are optimized to run on X86-64 technology, and take advantage of the superior CPU performance of the Intel Xeon 5000 sequence processors.

Connected with Gigabit Ethernet and/or InfiniBand, and accessing a common set of storage through high-performance SGI NAS platforms, these systems can support up to 160 CPU cores with almost 2 TFLOPS of raw performance in a single rack, with almost unlimited high-throughput scalability. But, most importantly, these systems come pre-configured and tested with Scali Manage™ and Altair® PBS Professional™ so that you can spend your time doing science rather than making computers work.

When integrated in a Scientific Workflow Solution for Computational Chemistry, Altix XE cluster systems allow high-throughput jobs to be run on systems which are explicitly tailored to the task, allowing high-performance jobs, those that require large memory, and any "overflow" high-throughput jobs to be run on the SGI Altix systems.

Data Solutions: High Performance Storage and Information Life Cycle Management

Data sets continue to explode in size, and have become a key differentiator for the increasing the impact of scientific research. High-performance data access is the limiting factor in some computational chemistry environments, and advanced versioning and data archiving techniques are required for regulatory approval.

The SGI family of storage platforms and data management environments are the underpinnings of the SGI Scientific Workflow Solution for Computational Chemistry. They allow multiple computer room systems to access a single copy of data in an efficient, high-performance environment and extend that single view of data directly to desktops. They also address an organization's data archiving needs with information lifecycle management software and high-performance, low-cost tape alternatives.

- SGI InfiniteStorage RAID arrays deliver leadership performance, price/performance and density for direct connect or shared file systems. SGI offers the industry's widest range of high-performance products with entry-level SAS and Fibrechannel RAID arrays to the InfiniteStorage 10000 which is a high-performance, cost effective alternative to tape archival and everything in between.
- SGI InfiniteStorage shared filesystem CXFS™ provides a new paradigm for data access that streamlines data flows, offers unparalleled shared data access over storage-area networks (SANs), and ensures that bandwidth limitations don't slow down critical applications.
- SGI InfiniteStorage NAS solutions extend that high-performance to networked users with up to 800MByte/S of performance over multiple GBit Ethernets, and an advanced GUI based management environment that gets you up and running in under 10 minutes.
- SGI Data Migration Facility (DMF) automates data migration for the highest possible capacity utilization, providing much faster access to data than manual data migration, and removing the human error potential from the data movement itself.

Accelerating Scientific Results

The SGI Scientific Workflow Solution for Computational Chemistry integrates the capabilities of SGI Altix systems, SGI Altix XE cluster systems and SGI InfiniteStorage capabilities into a single, unified computing environment. Workflows are accelerated because scientists can focus on their science instead of computer system access and data management. Furthermore, organizations are able to maximize overall productivity and minimize total cost of ownership because they are able to support the shifting needs of multiple research groups with a single, flexible solution.

Corporate Office
1200 Crittenden Lane
Mountain View, CA 94043
(650) 960-1980
www.sgi.com

North America +1 800.800.7441
Latin America +55 11.5185.2860
Europe +44 118.912.7500
Japan +81 3.5488.1811
Asia Pacific +1 650.933.3000

© 2006 SGI. All rights reserved. SGI, Altix, the SGI logo and the SGI cube are registered trademarks and NUMAflex, CSFS and The Source of Innovation and Discovery are trademarks of Silicon Graphics, Inc., in the U.S. and/or other countries worldwide. Linux is a registered trademark of Linus Torvalds in several countries. Intel, Itanium and Xeon are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries. All other trademarks mentioned herein are the property of their respective owners.

3970 [10.13.2006]

J15197