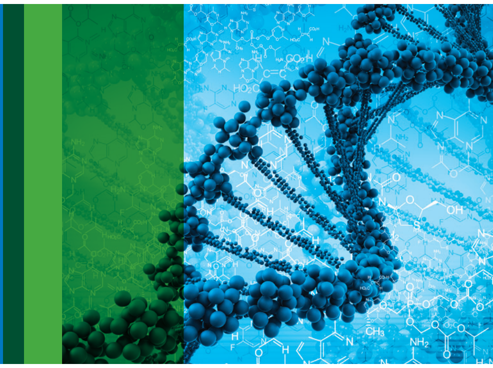
# Speed, Scale and Open Systems: SGI® Powers Genomics Research

## Genomics Research Accelerates with Purpose-Built SGI Technical Computing Solutions

In the dynamic world of genomics, there is one constant – the explosion of data generated in sequencing work. These fast-growing datasets demand new and more efficient approaches to quickly access, process and analyze results. SGI, with a long history in technical computing and providing computational solutions for life and chemical sciences, has kept pace with these trends by offering the most scalable compute and storage platforms in the industry.



Sequencing centers today are capable of generating the nucleic bases of thousands to millions of DNA fragments, resulting in the creation of many terabytes of data in a single day. Handling the sheer volume of this data influx is a challenge in itself, but this problem is compounded by the need to manage this dataflow through the demands of pre- and post-processing analysis. Together, this creates an extremely daunting technical computing and data management challenge.

The singular focus of SGI is technical computing leadership. We offer the technology needed to address these challenges, most notably the SGI® UV™ shared memory server platform and the SGI® InfiniteStorage™ ecosystem. These solutions share key attributes to accelerate success:

**Speed:** Fastest time to results

**Scale:** Industry-leading scale in capability, not just capacity

**Simplicity:** Open standards, familiar interfaces, factory integrated

## SGI Solution: Speeding the Entire Workflow, from Concept to Breakthrough Innovation and Discovery

### Overview

SGI, a leader in the life sciences community for more than 20 years, delivers computational solutions for life and chemical sciences discovery research organizations in pharmaceutical, chemical and biotechnology, as well as in academic and national labs. This background, along with our industry partnerships, positions SGI to make a real difference in delivering genomics solutions that are not only proven and productive, but also revolutionary, in their ability to drive breakthrough discovery.

**SGI Genomics Users Across the Globe**

Hundreds of genomics researchers enjoy the benefits and unrivaled capability of technical computing solutions from SGI. Already, users at life sciences institutions such as The Genome Analysis Center (UK), University of Minnesota (USA), Agency for Science, Technology and Research (Singapore), and Catalonia Supercomputing Centre (Spain), as well as major pharmaceutical companies, are putting SGI compute and storage technology to work.

Visit http://www.sgi.com/solutions/genomics/ for more information.

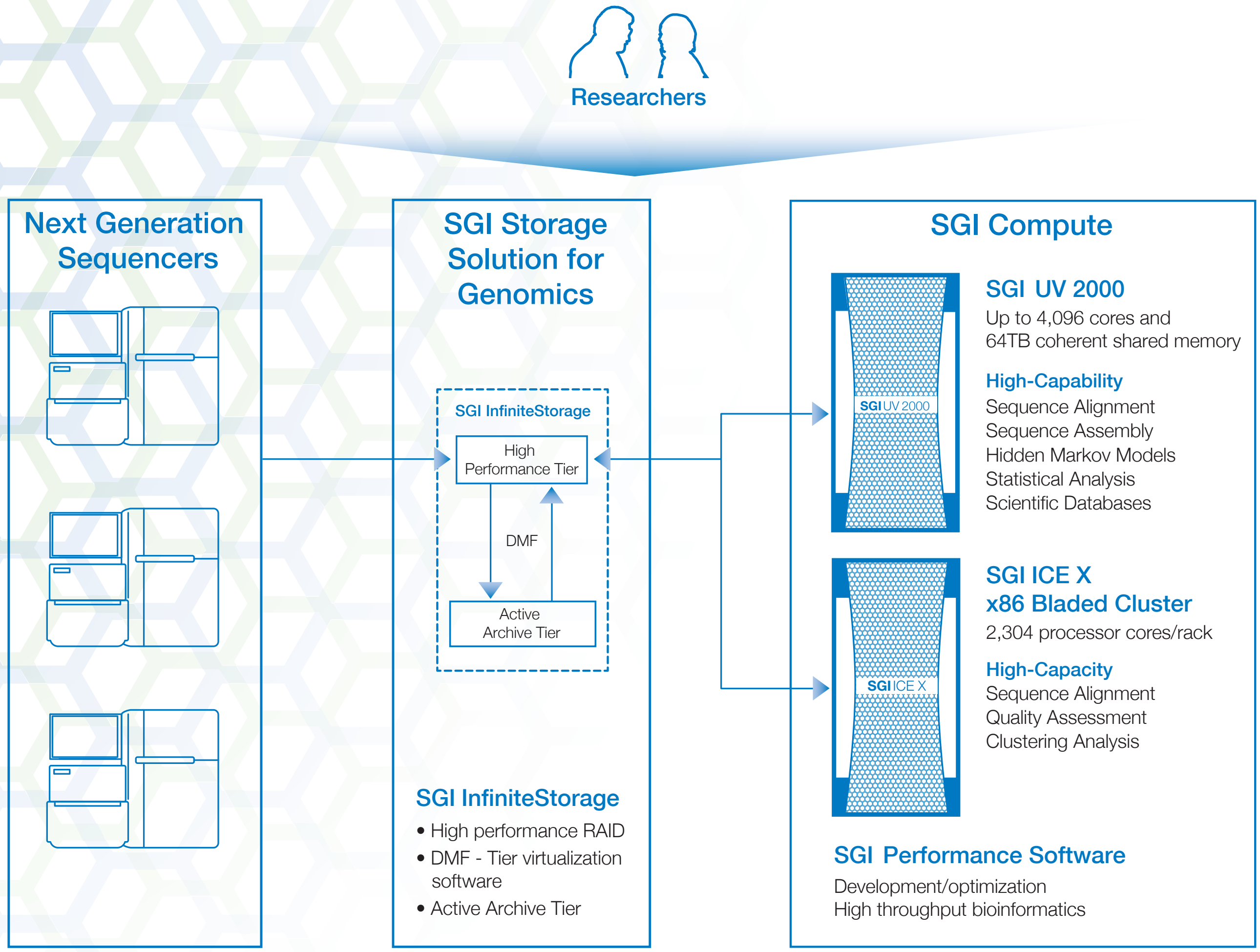


Figure 1. SGI solutions supporting a genomics workflow, taking terabytes of data and moving it through pre- and post-processing, keeping huge volumes of data immediately accessible to researchers.

## SGI UV Shared Memory Compute Platform

SGI UV is the largest capacity server available today as measured in number of processor cores and memory addressable in a Single System Image (a single instance of the operating system). The SGI UV system tops out at 2,048 processor cores and 64TB of memory – the maximum supported by the Intel® Xeon® processor E5 4600 product family. This large memory address space accommodates very large datasets in memory, where a variety of applications can then access the data without having to wait for relatively slow disk access. Applications can access data thousands of times faster using memory versus standard hard drives, providing the compute capability required for the most demanding applications such as sequence assembly jobs using Velvet, SOAPdenovo and ABySS. This scalability also provides the horsepower needed to address the increasing requirement to integrate genome data with data from other sources such as mass spectrometry, microscopy and medical imaging.

The SGI UV 2000 model offers up to 512 cores and 16TB of memory in a cabinet. A full-scale SGI UV 2000 is comprised of four racks and totals 2,048 cores and 64TB memory in one instance of the operating system.

In addition to its industry-leading system scalability, SGI UV also features an accelerated Message Passing Interface (MPI) and runs standard, off-the-shelf Linux®, so that x86 software can run unmodified.

Aside from performance, there are a couple of other important benefits. The first comes from the ease of code development and optimization that is inherent in a system that essentially operates just like a super powerful workstation. Another is the efficiency and ease of management that come from being able to handle entire computational workflows — including pre- and post-processing analysis — on one platform, without moving data.

## SGI ICE™ and SGI Rackable™ Cluster Solutions

The SGI Rackable cluster and SGI ICE bladed cluster product lines are purpose-built for technical computing. They deliver top performance and superior configuration flexibility in cost-effective solutions that are easy to build and deploy. These cluster solutions are ideal for high-throughput sequence comparisons and alignments using codes such as BLAST, FASTA and HMMER, where thousands of unknown query sequences are searched against multiple databases.

**With SGI solutions**, researchers gain a competitive advantage. Larger complex modeling and analyses can be completed in record time, with modeling done at higher theory levels. SGI solutions shrink time-to-market and reduce costs by providing:

- Scaling superiority for all model sizes and theory levels
- Flexibility to handle large and small simulations on a single server
- Fastest performance with computational biology applications
- Ease of management
- Industry standards

As industry leader in technical computing, SGI offers the premier software environment for high-performance science:

- Complete, integrated environment from deskside to supercomputer
- The highest levels of performance and scalability with full application compatibility
- SGI software and services that improve the productivity of users, developers and system administrators
- Extensible, based on open standards

## Managing the Data Flow

A key element to managing genomics workflow is the ability to quickly ingest, move and manage massive amounts of data. In typical workflows, the data generated averages around 6TB per day for each sequencer. This means that being able to have immediate "online" access to a huge amount of data, and rapidly route it into and out of post-processing engines, all in a way that keeps costs at a minimum, are vitally important.

With a comprehensive ecosystem of scalable storage products, SGI InfiniteStorage solutions have long been at the leading edge in high throughput and large volume data management. This includes shared file systems and scalable RAID platforms needed for this type of dataflow.

However, in most workflows, "transactional storage," the high-speed arrays that are used as the target of sequencing engines, is only practical for immediate access and short-term use. They are expensive and consume too much power and datacenter space to be used for long term, high volume data storage. But moving this data off of transactional storage to traditional archive environments, such as tape libraries, is far too slow and cumbersome for managing the rapid throughput required to move dozens of terabytes of data daily.

SGI offers storage solutions such as ArcFiniti, an active archive platform that directly targets this problem. As a network-accessible, disk-based, active archiver, ArcFiniti brings together the best of SGI on-line and off-line storage technology to provide a true active archive that has cost characteristics similar to tape, but with the high throughput and data protection demanded by genomics workflows. Each enclosure can have up to 1.4PB of usable active archive, and provides up to six 10GigE network ports to enable over 30TB of data movement per day.

With features such as proactive health monitoring of hardware and data integrity checking, ArcFiniti ensures that archival data is accessible and valid for a dramatically longer lifecycle than traditional disk or tape solutions. If ArcFiniti detects any potential disk drive problems, it proactively migrates data and verifies data integrity while alerting administrators to replace the faulty part. This feature dramatically reduces the need to take the system offline for costly RAID rebuilds due to disk errors.

## Virtualized Tiers

As a fully integrated solution, ArcFiniti is designed to be easy to deploy and easy to use. Users see the entire system as one unified online archive pool. But to provide data protection and performance, ArcFiniti virtualizes a high-performance primary cache with an archive tier based upon the SGI award-winning MAID technology. This virtualization is managed automatically in the background by an archive policy engine. All files are always in an available online state to users, with ArcFiniti ensuring that archive content is protected for long-term retention on the most cost-effective storage tier.

ArcFiniti is not only power efficient but also enables storage consolidation with minimum datacenter space.

## SGI Software: Performance and Productivity, Plus High-Throughput Computing Bioinformatics Wrapper

SGI system management software supercharges the capabilities of the operating system to maximize performance on SGI computer systems, while enabling researchers to run jobs when resources become available based upon an array of parameters that datacenter managers set up.

For genomics researchers, SGI has an additional domain-specific solution. Algorithms like BLAST and HMMER face scalability challenges that impede performance. Also, start-up overhead associated with launching jobs can become very significant when multiplied by thousands of jobs. SGI application engineers developed a wrapper program called the HTC (High Throughput Computing) Wrapper Program for Bioinformatics. HTC is a single binary download, available for free, that transparently reads all the inputs, load balances the jobs, and then submits them to maximize system utilization and efficiency. HTC has been shown to demonstrate significantly higher performance and scalability than a general batch scheduler for processing thousands of query sequences, accelerating the throughput of any user-supplied script as well as a variety of bioinformatics codes already available such as BLAST, FASTA, ClustalW, HMMER and Wise2.

## SGI Genomics Solutions are Proven, Productive

### SGI Application Expertise Assures Optimal Solutions

SGI application engineers are constantly testing and manipulating the most popular genomics codes in order to assure their optimization for performance and productivity, and that solution configuration guidance can be provided for the best workflow support in a given organization. Here are but two of the many examples:

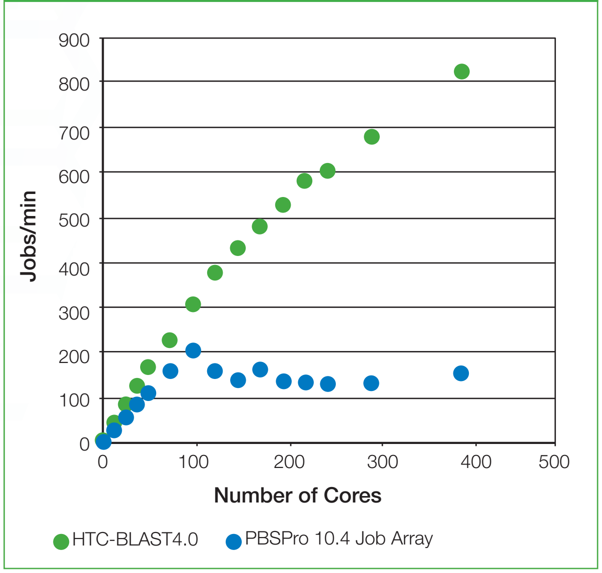
**1) SGI HTC-BLAST** reduces the execution time from over 60 hours on a single core to 18 minutes on 384 cores of an SGI UV system, yielding results for nearly 800,000 BLASTP queries per day. This corresponds to an increase in throughput rate from three jobs per minute on a single core to almost 550 jobs per minute on 384 cores.

**2) SGI Application Expertise:** The following graph shows how SGI application performance expertise, spanning various cluster solutions and our shared memory servers, enables us to provide trusted guidance on the optimal configuration to deploy for a specific set of user workloads.

## Conclusion

The SGI genomics solution can drive huge gains in time-to-solution for compute- and data-intensive workloads. This performance and versatility means great TCO/ROI for users and offers entirely new possibilities for breakthrough discovery. Let us show you how these solutions can work for you.

**HTC-BLAST 4.0 vs. PBSPro 10.4 Job Array with NCBI-BLAST 2.2.23+ (blastp) on 3.47 GHz/12MB SGI ICE 8400**

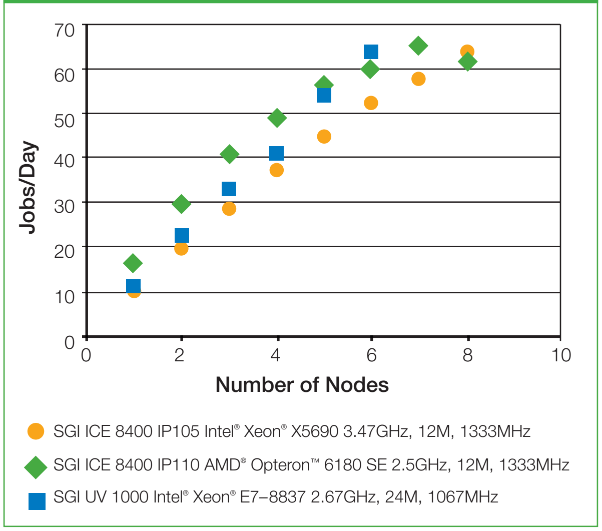


**Queries:** 10,000 sequences from RefSeq protein database (3,221,557 residues in 10,000 sequences)

**Database:** Non-Redundant Protein Database (1,336,756,300 residues in 3,878,509 sequences)

- HTC is 1.5x to 5.4x faster than using PBSPro Job Arrays.
- Relative speed-up of HTC increases with the number of cores.

**FASTX 35.4.10**
**1 Bovine Sequence vs NR Database**



**Global Sales and Support**: sgi.com/global