



TGen/ASU Advances Genomic Research with High-Performance Cluster based on Intel® Processors and Linux*

SOLUTION SUMMARY

Challenge	The Translational Genomics Research Institute* (TGen) is at the cutting edge of unraveling the genetic components of common and complex diseases. Its mission is to combine genomic research with computational biology and translate results into practical interventions for diseases such as melanoma, diabetes, and Alzheimer's. Genomics research is highly dependent on massive computational capabilities. TGen needed to construct a supercomputer that would allow it to cost-effectively scale to thousands of processors.
Solution	TGen and one of its principal academic partners, Arizona State University (ASU), chose an Intel® architecture-based Linux* cluster rather than a RISC-based architecture because of its superior price/performance. ASU purchased a 512-node cluster of IBM* eServer xSeries* 1350 servers powered by 1,024 Intel® Xeon™ processors and running the Red Hat* Linux operating system.
Business value	Intel architecture is less expensive than a RISC-based architecture, allowing TGen and its partners to benefit from more computing power per research dollar. With the money saved on Intel-based clusters, ASU was able to purchase more compute nodes to more rapidly solve more complex genomics problems. This allows researchers to more rapidly bring treatments to market. An Intel-based supercomputer also lowers administrative and management costs.
Supercomputer	512-node cluster of IBM eServer xSeries 1350 servers powered by 1,024 Intel Xeon processors
Operating system	Red Hat Linux version 7.5
Applications	Gene expression analysis programs such as BLAST*, PAGE*, and BLAT* Phylogenetic analysis programs such as Tree-Puzzle* and FastDNAMl* Computational chemistry and molecular modeling applications such as AMBER and CHARMM Mathematical analysis and statistical applications such as GAUSS* Molecular graphics applications such as PovRay*
Storage	10 TB IBM FAST700 Storage Area Network (SAN)

Business Challenge

ENLISTING SUPERCOMPUTERS TO BREAK DOWN THE HUMAN GENOME

As a result of the Human Genome Project, genomic research is now entering an era where emerging data will enable investigators to ultimately unravel the genetic components of common and complex diseases. The much-anticipated complete sequence of the human genome, coupled with the emergence of the sequences of other animal, plant, and microbial genomes, now provides scientists with an unparalleled resource to address biological and medical questions.

The Translational Genomics Research Institute (TGen), founded in July 2002, seeks to discover the genetic changes underlying a variety of human diseases and deliver these discoveries to the patient bedside as improved healthcare interventions. This new chapter in genetic research is called "translational research" because it combines the emerging fields of genome-wide array technology and computational biology to provide the data and tools necessary to identify the genes that play a role in hereditary susceptibility to disease.

“The use of standards-based software and off-the-shelf processors has enabled ASU to reduce the operational support costs of the facility.”

Dr. William Lewis
Chief Information Officer
Arizona State University

Intel worked closely with the Arizona State Government and Arizona State University (ASU) in bringing TGen to the Phoenix area, where it has become the pillar institute for a growing life sciences industry. An Intel® Solution Services team, ASU, and TGen worked together to specify TGen’s computing needs and design the super-

computer architecture that was submitted to manufacturers of high-performance Intel-based computer systems for proposals. IBM provided the winning design with a 512-node cluster of IBM® eServer xSeries® 1350 servers powered by 1,024 Intel® Xeon™ processors and running the Linux® operating system.

“The computational side of genomics is incredibly important,” explains Richard Love, chief operating officer of TGen. “Making sense of 30,000 genes requires sophisticated pattern recognition software, which drives you to supercomputing. Having ASU, Intel and IBM on board gives us confidence that we’re getting the state of the art in high-performance computing.”

INTEL® ARCHITECTURE GIVES SCIENTISTS MORE POWER PER DOLLAR

The task of building a complete supercomputer infrastructure from the ground up fell to Dr. Edward Suh, Chief Information Officer of TGen, and Dr. William Lewis, Chief Information Officer of ASU. Dr. Suh was familiar with Intel-based cluster computing from his previous position at the National Institutes of Health (NIH).

“The Intel architecture is less expensive than RISC-based architectures traditionally used in building supercomputers,” Suh says. “And the performance is about the same for many of our applications. We benchmarked a RISC-based system against a comparably configured Intel-based cluster and found that several applications actually ran faster on the Intel system.”

With the money life science researchers save on Intel architecture-based clusters, they can buy more compute nodes and grind through problems faster. “If we had purchased a RISC-based supercomputer, we would have ended up with a much smaller system than the 512-node system we have—perhaps only a 100-node system for the same amount of money,” Suh says. ASU purchased and houses the joint-use cluster in support of TGen’s and ASU’s research mission.

Suh knew that TGen researchers would need the unparalleled volume economics of the Intel architecture to scale up to thousands of processors at a reasonable cost. An earlier test had demonstrated to Suh that the number of processors in a cluster reduces processing time almost linearly. Suh used for his test a complex genomics problem involving 600 genes. The problem was to understand how these genes interacted and involved testing all possible interaction combinations. Suh calculated that a single-CPU 2.4 GHz computer would solve the problem in one to two years. However, if he used 100 CPUs, he could reduce the computational time nearly 75-fold (some of the processing power is lost to overhead). “We would complete the 600-gene analysis in about one to two weeks using a 100-CPU system,” he says.

For the kinds of problems TGen researchers were tackling, thousands of processors would be required. The Intel architecture has allowed TGen/ASU to build one of the biggest supercomputers in the world.

e-Business Solution

INTEL® XEON™ PROCESSOR BUILDING BLOCK

A team of seven people—three at TGen and four at ASU—worked with the Intel Solution Services team to design the supercomputer. “Intel’s assistance in both system specification and site planning was invaluable in helping TGen and ASU establish a world-class computational facility,” Dr. Lewis says.

The TGen/ASU supercomputer employs a typical Beowulf cluster architecture, with 512 two-way nodes interconnected by Gigabit Ethernet. Within the 512-node cluster, 128 nodes are connected by Myricom Myrinet®, a high-performance, packet-communication and switching technology that is widely used to interconnect clusters of workstations, PCs, servers, or single-board computers. A Beowulf cluster is any high-performance massively parallel computer built primarily from commodity off-the-shelf hardware components, running a free-software operating system such as Linux or FreeBSD®, and interconnected by a private high-speed network. The Beowulf architecture is highly effective, and cost-effective, at running massive life and earth science applications.

“This kind of performance allows researchers to run analyses and receive answers in hours or days versus months or years.”

Dr. Edward Suh
Chief Information Officer
TGen

“Intel-based systems are far less expensive to scale than RISC-based systems, allowing us to stretch our research budgets and accomplish more science on a smaller budget.”

Dr. Edward Suh
Chief Information Officer
TGen

TGen and ASU selected the 2.4 GHz Intel Xeon processor as the basic supercomputer building block. “At the time we designed our computer, the Intel Xeon processor was the most established and cost-effective CPU on the market,” Suh says. “It’s a very powerful processor for cost-effectively scaling life sciences applications.” Each two-way node shares 2 GB of

RAM and has 36 GB of local disk of its own. Eight additional nodes are dedicated to providing I/O between the 512 compute nodes and a SAN-based RAID disk storage subsystem with 10 TB of storage.

The Intel Xeon processor is Intel’s cost-effective DP processor that features Intel® Hyper-Threading Technology, the Intel® NetBurst™ microarchitecture, and 512KB of Level 2 advanced transfer cache. Hyper-Threading Technology enables multithreaded software to execute threads in parallel, reducing overall processing time. The NetBurst microarchitecture enables higher transaction rates and faster response times. A 533 MHz system bus provides higher throughput when accessing memory and I/O devices for improved server headroom and scalability.

COST-EFFECTIVE PERFORMANCE, SCALABILITY

The TGen/ASU supercomputer is one of the 50 largest supercomputers in the world today based on benchmarking by the Top 500 Supercomputers Site (www.top500.org). Based on the results from a standard Linpack* benchmark, the TGen/ASU supercomputer ranks among the 50 fastest in the world, with performance of 1,750 GFLOPS (thousands of floating-point operations per second).

“This kind of performance allows researchers to run analyses using a variety of life sciences applications and receive answers in hours or days versus months or years,” Suh says. Some of the TGen applications are floating-point-intensive, others are memory-intensive, but all share the common characteristic of very large data sets. One of TGen’s primary applications is BLAST*, a set of programs for finding similarity between a query protein or deoxyribonucleic acid (DNA) sequence and a sequence database.

The TGen/ASU supercomputer is currently supporting a half dozen research programs, but eventually there will be hundreds of users on the system. Suh plans on partitioning the supercomputer into 10 or 15 sub-clusters dedicated to running specific applications and data sets

directed at solving specific genomic problems. Suh expects the current cluster to take care of TGen’s needs for the next three to four years. However, when the time comes to expand, the modular, cost-effective Intel architecture will make growth easy.

“As we grow, we have two scale-out strategies,” Suh says. “We can add additional nodes or we can upgrade the processors, memory, and interconnect technology in our existing system. Intel-based systems are far less expensive to scale than RISC-based systems, allowing us to stretch our research budgets and accomplish more science on fewer dollars.”

An Intel architecture-based supercomputer is also less expensive to care for and maintain than a RISC-based system. There’s a wealth of programmers and support technicians familiar with the Intel architecture; parts are less expensive; and there are more, and therefore more cost-effective, server management and administrative tools available. “The use of standards-based software and off-the-shelf processors has enabled ASU to reduce the operational support costs of the facility,” Lewis says.

DOWN THE ROAD: INTEL® ITANIUM® 2 PROCESSOR PROVIDES MORE MEMORY

TGen and ASU are also eager to evaluate clusters based on Intel® Itanium® 2 processors. “A 64-bit system will allow us to address problems with larger memory,” Suh says. Offering performance typically associated with RISC-based processors, the Intel Itanium architecture includes a 64-bit address base, the ability to work with larger numbers and data sets, and higher precision in performing floating-point operations faster. The Itanium architecture is designed to meet the computational needs of those working with integer and floating-point applications, as well as people working with large-scale data sets.

INTEL BRINGS BROAD RESOURCES TO LIFE SCIENCES

Intel brought more than just processor technology to the TGen table. Intel CEO Craig Barrett chairs the Dean’s Council at ASU’s Fulton School of Engineering and was consulted early on by then-Arizona Governor Jane Dee Hull for help in recruiting TGen to the state. Intel Solution Services, Intel’s professional services

“In life sciences, computer speed has a direct impact on scientific discovery, and our Intel-based cluster promises to help us make rapid breakthroughs that will positively impact human health.”

Richard Love
Chief Operating Officer
TGen

“The Intel architecture is less expensive than RISC-based architectures, and the performance is about the same for many of our applications.”

Dr. Edward Suh
Chief Information Officer
TGen

organization, provided design consultation services in specifying and requesting proposals for the supercomputer. Intel consultants also assisted in the design of the ASU facility where the computer resides.

TGen is using Intel® C/C++ compilers for development work because of their superior perfor-

mance. “The Intel compilers are about 10 to 15 percent faster than others we tested,” Suh says. TGen also uses open-source GNU compilers and IBM compilers. Intel Solution Services consultants provided the TGen developers with training on how to optimize their life sciences applications for Intel architecture.

“Intel’s involvement was instrumental in getting TGen located in Arizona and getting our supercomputer up and running in a record three months,” Love says. “Intel is such a strong, well-respected firm. We are confident that we have the best supercomputer technology available for solving genomic problems. The faster we can work through thousands of gene analyses, the faster we can come up with new treatment approaches for debilitating diseases. In life sciences, computer speed has a direct impact on scientific discovery, and our Intel-based cluster promises to help us make rapid breakthroughs that will positively impact human health.”

LESSONS LEARNED

- **Use cost-effective Intel architecture-based clusters to tackle bioscience problems.** TGen/ASU selected a Beowulf cluster based on 512 two-way Intel® Xeon™ processor-based nodes. Intel architecture allows TGen to scale to thousands of processors much more cost-effectively than they could using a RISC-based architecture. More processors can yield scientific breakthroughs sooner.
- **Optimize clusters to applications.** Life sciences applications have a wide range of application characteristics: some are floating-point-intensive, some memory-intensive, some both. A solid understanding of application characteristics will determine which processors to use in building a cluster. Intel offers a range of solutions from the dual-processing Intel Xeon processor to the 4-way Intel Xeon processor MP and the Intel® Itanium® 2 processor.
- **Establish a support team.** Have administrative and programming teams trained and ready to go before deployment is complete, Suh says. The TGen/ASU support team of 13 people had received training from IBM and Intel before the supercomputer went live, allowing researchers to take immediate advantage of the new computing resource.

Intel works with the world’s largest community of technology leaders and solution providers—from software and hardware to systems integration and services companies—that are all using Intel® products, technologies and services with a common goal of providing better, more agile, cost-effective business solutions for you.

Find out more about a business solution that is right for your company by contacting your Intel representative, or visit the Intel® Business Computing Web site at: intel.com/ebusiness or its industry solutions specific sites: intel.com/go/retail, intel.com/go/manufacturing, intel.com/go/digitalmedia, intel.com/go/finance, intel.com/go/telco, intel.com/go/hpc, intel.com/go/energy.



Information in this document is provided in connection with Intel® products. Except as provided in Intel’s terms and conditions of sale for such products, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY RELATING TO SALE AND/OR USE OF INTEL PRODUCTS, INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT, OR OTHER INTELLECTUAL PROPERTY RIGHT. Intel products are not intended for use in medical, life-saving, life-sustaining, critical control or safety systems, or in nuclear facility applications.

Intel may make changes to specifications, product descriptions, and plans at any time, without notice.

Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance. Buyers should consult other sources of information to evaluate the performance of systems or components they are considering purchasing. For more information on performance tests and on the performance of Intel products, reference www.intel.com/procs/perf/limits.htm, or call (U.S.) 1-800-628-8686 or 1-916-356-3104.

Intel, the Intel logo, Intel Xeon, and Itanium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.

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