SOLUTION BRIEF

UNIVERSITY OF HAWAII'S PAPAYA GENOME SEQUENCING PROJECT

Papaya Genome Sequencing Project Opens Vast Array of Agricultural, Scientific and Culinary Opportunities

"We knew that we needed a big and powerful system to do what we call 'assembly' in which we input around 5GB of data into the program and this generates around 150GB of output that we need to analyze, so it's very good that we have this Altix."

–Alexandre Dionne-Laporte, bioinformatic software engineer, University of Hawaii at Manoa

The fifth largest crop in Hawaii, papaya was almost wiped out by the papaya ringspot virus a decade ago. Genetically modifying the papaya to create a ringspot-resistant transgenic or "hermaphrodite" variety saved the papaya from extinction on the islands. The papaya is now the first fruit – and the first transgenic crop – to be sequenced, helping to pave the way for international export of genetically modified fruits and vegetables as well as opening up a surprising number of fields of scientific research.

Two years ago, researchers at the University of Hawaii (UH), Dr. Takashi Sugimura, Maui Super Computer Center (MHPCC), Dr. Stan Saiki, Pacific Telehealth and Technology Hui, Dr. Ray Ming, University of Illinois at Urbana-Champaign then at Hawaii Agricultural Research Center (HARC), Dr. Paul Moore, USDA Pacific Basin Agricultural Research Center together with Dr. Lei Wang, Nankai University began the Hawaii Papaya Genome Project to map papaya genes to improve the efficiency of agricultural cultivation, and to discover new applications for one of the most important edible fruit crops of tropical and subtropical regions. Papaya is also used in a wide range of medical, biotechnological and cosmetic applications – many based on the enzyme papain (also used as a meat tenderizer) – that can be further developed once a complete picture of the genomic sequence has been drawn.

The Genome project will allow for the identification of agronomically important genes, which is expected to have a wide-reaching agricultural and economic impact. The sequence

UNIVERSITY OF HAWAII'S PAPAYA GENOME SEQUENCING PROJECT

"With the SGI Altix, we chose it because of the configuration of the memory system, how quickly we can use the random memory, and also the scalability of the system, and of course, the pricing. We also have an interactive Web site, Gbrowse, for our national and international collaborators on the project. PipelineFX had just come out with Qube!, which handles distribution tasks, and that was a tremendous help. PipelineFX, right here in Hawaii, gave us a package that I thought no one would be able to bid on."

–Dr. Maqsudul Alam, University of Hawaii at Manoa

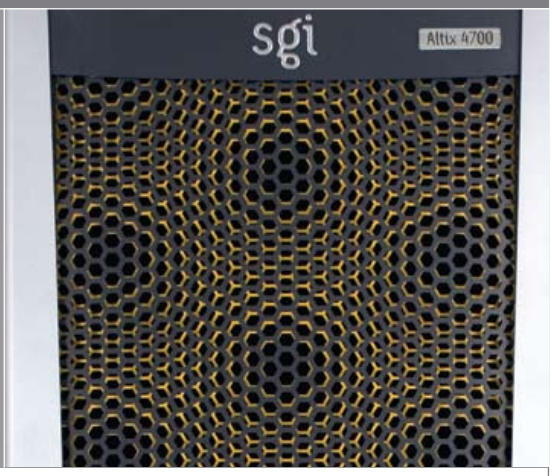
information may help develop papayas that only produce hermaphroditic fruits, which are preferred on the female. This will result in reducing the costs of establishing papaya fields and require less labor in differentiating hermaphrodite from female plants. Scientists will also be able to study sex chromosome evolution, because the papaya actually displays biological features of a "living fossil" among plants and provides an exceptional opportunity to study sex chromosome evolution and sex determination in a variety of organisms. Papaya offers several features that make it highly suitable from a scientific point of view for a project of this magnitude in Hawaii, including a comparatively small genome. Genomes are measured in what are called "base pairs" and there are 372 million base pairs in the papaya genome. Oddly enough, given the size of the fruit, the papaya genome is approximately 14% smaller than the rice genome.

In the 1980 and 90s, UH alum Dr. Dennis Gonsalves while at Cornell University led a research team of Drs. Jerry Slightom, Richard Manshardt, and Maureen Fitch in a research effort that resulted in the creation of the first transgenic papaya. The transgenic papaya is currently raised throughout Hawaii because it is resistant to the ringspot virus while not losing any fruit-bearing capabilities or flavor. The sequencing of the genome is helping investigators in their efforts to obtain approval of Hawaii's transgenic papaya in Japan, a very important export market for Hawaii's papaya. Undoubtedly, the sequencing of the genome will aid efforts in deregulating the transgenic papaya in places as European countries, who may require detailed information on the position of the transgenes in the papaya genome.

SGI and PipelineFX Power University

"Dr. Alam and his team put a lot of effort into building a Web front-end for Blast and ClustalW and for the UH research community. With the open APIs that come with Qube!, they could pre-load all the standard databases, up to the 92GB of memory they have on the SGI Altix server, so that when you do a Blast search, you're literally searching out of memory and it is super fast."

–Richard Lewis, VP of Sales, PipelineFX



Meeting the Computational Requirements

While the 372 million base pairs that make up the papaya's nine chromosomes is not a monumentally huge project (humans have 23 chromosomes and billions of base pairs), sequencing any genome requires a vast amount of computation. This led the Center to seek the fastest, largest computer that their budget would allow, one that was actually capable of reliably running the 2-year-plus project from the beginning. Dr. Maqsudul Alam, who heads the papaya genome project, wanted a computer configuration that would also service the university's fast-growing bioinformatic research community as a whole. Dr. Alam also needed to replace the researchers' dependence on the free National Center for Biotechnology Information (NCBI) multi-thousand-processor cluster on the U.S. mainland.

NCBI is a shared resource used by many national researchers to access gene and protein sequence information from a number of popular open-source Life Sciences databases, such as Blast, ClustalW and Fasta. Since NCBI is a free resource, turnaround time and speed of searches can be extremely slow because the memory is not shared across the clusters.

In addition to processing power and speed, Dr. Alam looked for a backend distributed processing solution that was scalable, had open API's to develop Web-based interfaces to applications only available as command line tools, and allowed the development of custom processes such as pre-loading standard databases into memory for faster searches. Another major consideration in their choice of hardware and software was that the Hawaii Papaya Genome Project was going to generate a tremendous amount of data – a typical papaya base pairs assembly run can use as much as 40GB of memory and 150GB of disk space and take from 7 to 10 days to calculate – and, with future plans of doing even larger sequencing projects, the researchers would require very large, and highly scalable, disk storage.

All four design challenges – a powerful computer for the specific project, a server and storage system that could also function as a local department resource, the ability to write custom Web-based front end interfaces, and scalable memory and scalable storage – were met by the combination of the SGI® Altix® 350 computer system, SGI® InfiniteStorage TP9300 Fibre Channel storage, which has been upgraded over the course of the project from 4TB to over 16TB, and PipelineFX™ Qube!™ render management and batch-queuing software. The complete system was integrated and sold through CORESystems Hawaii, the SGI reseller in Hawaii that has extensive hardware and middleware experience with Altix servers and PipelineFX Qube! software. This met the fifth requirement: a Hawaii-based firm with solid mainland connections.

"Hawaii has its own pros and cons," said Dr. Alam. "If something goes wrong, it's a five and a half hour flight from the mainland, minimum. Whenever we choose a big instrumentation addition, we always have to think about who is a local partner that we can trust and a local partner that has extremely good relations with a mainland partner. With the SGI Altix, we chose it because of the configuration of the memory system, how quickly we can use the random memory, and also the scalability of the system, and of course, the pricing. We also have an interactive Web site, Gbrowse, for our national and international collaborators on the project. PipelineFX had just come out with Qube!, which handles distribution tasks, and that was a tremendous help. PipelineFX, right here in Hawaii, gave us a package that I thought no one would be able to bid on."

The SGI Altix 350 system, with 92GB memory and 32 Intel® Itanium® 2 processors (upgraded last year from the original 16 processors), has the unique SGI CCNUMA architecture that allows very large memory to be shared across all the processors. The Center personnel wrote their own software that, every night, downloaded all the versions of the genomic databases, and the most common were uploaded into the Altix system's shared memory for

immediate access. To better protect the SGI Altix from Hawaii's tropical storms and power outages, the system was moved from the UH Manoa campus to the Maui High Performance Computing Center (MHPCC) and is accessed over the WAN.

"We knew that we needed a big and powerful system to do what we call 'assembly' in which we input around 5GB of data into the program and this generates around 150GB of output that we need to analyze, so it's very good that we have this Altix," said Alexandre Dionne-Laporte, bioinformatic software engineer, University of Hawaii at Manoa. "We needed a system that has a lot of processing power but still a system that can provide us a lot of memory – shared memory – for programs to run efficiently. We also needed a system that was stable and that could run jobs for weeks without stopping. We selected SGI for the price-performance, the shared memory, scalable storage and because it is a single system machine, as well as for the local service and availability in Hawaii. We are very happy with this system."

Enabling Scientists to Do Research, Not Computer Programming

For the processing power, the 32 CPUs give the team high throughput to their computing needs as they can run highly multi-threaded applications very efficiently and concurrently. The huge amount of shared memory is also nice, since they have application that can use up to 40GB of memory in a single run. The team also uses a SGI port of ClustalW to run in a parallel fashion on the Altix. PipelineFX Qube!, used widely in digital media for game build engines and animation rendering, was uniquely set up as the Center's backend software environment. Qube! software is written with a very open API, C++, Pearl and Python, and everything that PipelineFX develops within their software -- all the source APIs -- are available directly to customers. The openness allows easy access to almost every aspect of the software, especially all the interfaces and hooks to key features and monitoring tools.

This allows the Center's researchers to write different interfaces on top of Qube!, which in turn allows them to share server and storage resources for multiple tasks. Now they can do their research using Qube! as a Blast, ClustalW and Fasta resource manager, as well as a very powerful genomic sequencing machine.

"Dr. Alam and his team put a lot of effort into building a Web front-end for Blast and ClustalW and for the UH research community," said Richard Lewis, VP of Sales, PipelineFX. "With the open APIs that come with Qube!, they could pre-load all the standard databases, up to the 92GB of memory they have on the SGI Altix server, so that when you do a Blast search, you're literally searching out of memory and it is super fast."

The processing power and reliability of the combined SGI and PipelineFX solution has kept the Hawaii Papaya Genome Project right on schedule. For example, Blasting a 550 base pair (bp) sequence against nucleotides (nt), 17,918,478,829 bp takes about five seconds to complete, which means Center researchers could run more than 16,000 Blasts per day. Nucleotides are nitrogen-containing molecules which link together to form strands of DNA and RNA; pairs of complementary nitrogenous bases interact to form each rung of DNA's double helix. Theses results reflect the papaya genome sequencing at 82% completion, with full completion expected before the end of the year.

The Hawaii Papaya Genome Project is a multi-institutional bioinformatics project. In Hawaii, participants include MHPCC, Hawaii Agricultural Research Center (HARC), Hawaii Papaya Industry Association, USDA Pacific Basin Agricultural Research Center, Pacific Telehealth and Technology Hui, and Hawaii Biotech. Nankai University in China is a major collaborator, as is Maryland Biotechnology Institute.

Papaya Genome Website at University of Hawaii
http://asgpb.mhpcc.hawaii.edu/papaya/

sgi®

Corporate Office
1140 E. Arques Avenue
Sunnyvale, CA 94085
(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

4022 [07.2007]
J15303