San Diego Supercomputer Center

Magazine of Cyberinfrastructure & Computational Science

ENVISION

VOL. 21, NO. 1

FALL 2005

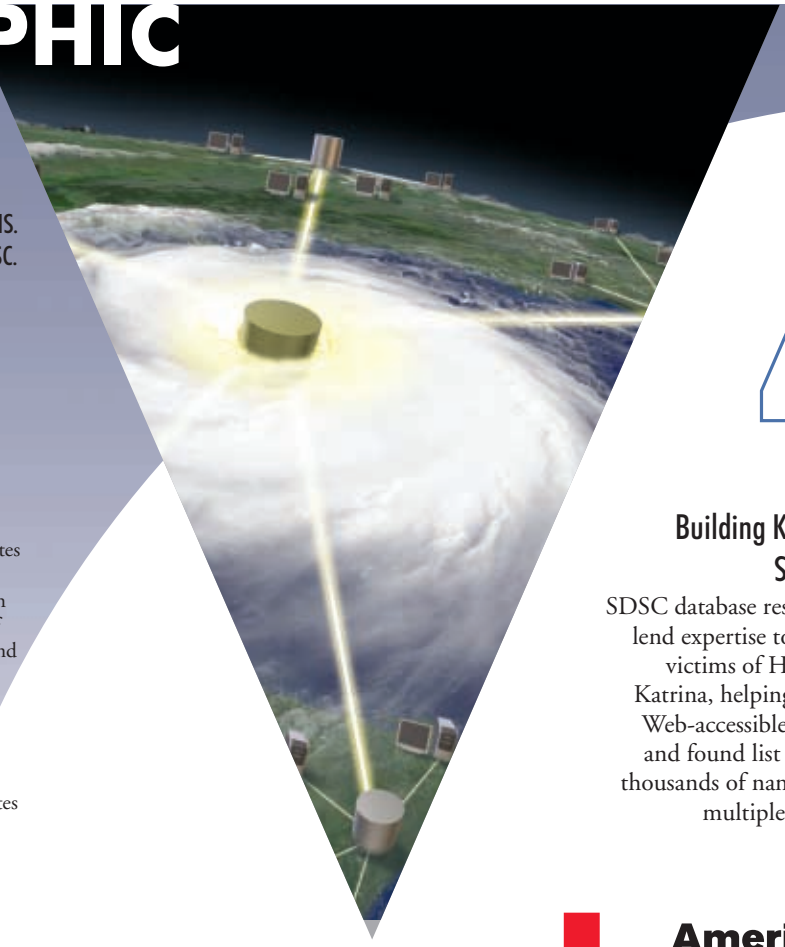
DATA *and* DISASTER

Building KATRINA'S Superlist

SDSC Data Experts Help Reunite Hurricane Katrina Evacuees

SDSC

COVER GRAPHIC



Weather satellite view of Hurricane Katrina, image courtesy NOAA/NESDIS. Cover graphic, Ben Tolo, SDSC.

THE SAN DIEGO SUPERCOMPUTER CENTER

In 2005, the San Diego Supercomputer Center, SDSC, celebrates two decades of enabling international science and engineering discoveries through advances in computational science and high performance computing. Continuing this legacy into the era of cyberinfrastructure, SDSC is a strategic resource to academia and industry, providing leadership in Data Cyberinfrastructure, particularly with respect to data curation, management, and preservation, data-oriented high-performance computing, and cyberinfrastructure-enabled science and engineering. Primarily funded by the NSF, SDSC is an organized research unit of the University of California, San Diego and one of the founding sites of NSF's TeraGrid. For more information see www.sdsc.edu.

SDSC INFORMATION

Fran Berman, *Director*

Vijay Samalam, *Executive Director*

San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive MC 0505
La Jolla, CA 92093-0505

Phone: 858-534-5000

Fax: 858-534-5152

info@sdsc.edu

www.sdsc.edu

Greg Lund, *Director of Communications*
greg@sdsc.edu
858-534-8314

ENVISION

enVision magazine is published by SDSC and presents leading-edge research in cyberinfrastructure and computational science.

ISSN 1521-5334

EDITOR: Paul Tooby

CONTRIBUTORS: Cassie Ferguson, Lynne Friedmann, Greg Lund, Paul Tooby, Ashley Wood

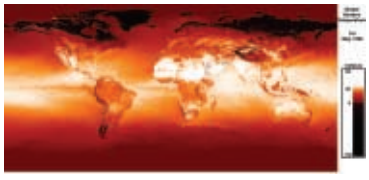
DESIGN: Beyond The Imagination Graphics

FROM THE DIRECTOR

FEATURES

THE BACK COVER

Long-term Records of Global Surface Temperature



From the Director

DATA *and* DISASTER

Over the last few years, the world has been both sobered and galvanized by the immense impact of large-scale natural disasters. The immediate and massive loss of life and property during the **Indian Ocean tsunami** and **Hurricane Katrina**, as well as the longer term impacts on the local, national, and global economy and ecosystems from these events will have repercussions for many, many years to come.

"SDSC staff worked with the Red Cross, the National Institute for Urban Search and Rescue, San Diego State University, Microsoft, and others to create an amalgamated 'missing and found' persons list with many thousands of names compiled from separate sources."

The interdisciplinary science, engineering, and information technology staff at SDSC work each day with researchers and educators supporting and serving users, and enabling advances and new discovery through Cyberinfrastructure. SDSC staff recently applied this experience to help the Red Cross amalgamate critical survivor data. Our dedicated and generous team worked around the clock to create one list that has become the central Red Cross resource for people trying to locate lost loved ones in the wake of Hurricane Katrina.

For SDSC, and our colleagues in the science, engineering, and technology communities, the opportunity to give back to the broader society at a time of real need using approaches, tools, and techniques pioneered for academic research, is immensely gratifying.

Katrina was not the first experience with disaster data for SDSC staff, who have also worked with tsunami data from the Indian Ocean and earthquake data from the Southern California Earthquake Center. In this issue of EnVision, we focus on ways in which SDSC's Data Cyberinfrastructure and the expertise developed from working with science and engineering communities can make a difference more broadly—through predicting, responding to, and evaluating the impact of large-scale natural disasters.

THREE PERSPECTIVES

For both survivors and family and friends, sifting through a rapidly increasing number of websites with survivor data to find a missing loved one (30+ sites when SDSC was brought in) is a time-consuming and onerous task. Using database technologies originally designed to coordinate multiple science and engineering data collections, SDSC staff worked with the Red Cross, the National Institute for Urban Search and Rescue, San Diego State University, Microsoft, and others to create an amalgamated "missing and found" persons list with many thousands of names compiled from separate sources. The technical challenges involved translation and coordination of multiple data formats, data cleaning and "de-duping" (removing

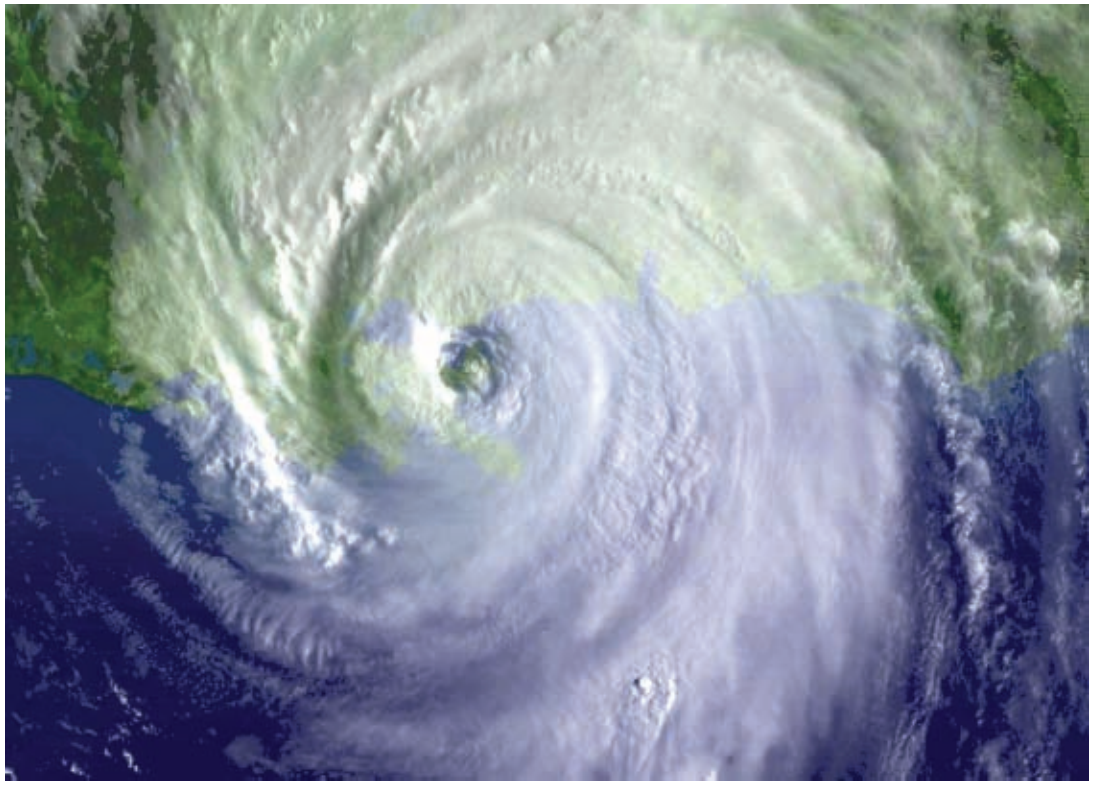
by Dr. Francine Berman, SDSC Director

duplicates), the addition of a "fuzzy" search quality which allows close but inexact queries to obtain useful information, etc. The resulting KatrinaSafe.com website provides one-stop shopping through a "meta-list" of survivor data. More about SDSC's Katrina efforts can be found in the cover story on page 4.

Katrina data was collected and used in real-time during the first intense weeks following the hurricane. Following the December 2004 Indian Ocean tsunami, more than 20 National Science Foundation funded scientific reconnaissance teams, supported by the SDSC-based Network for Earthquake Engineering Simulation Cyberinfrastructure Center (NEESit), went to work in Asia, gathering voluminous data from the tsunami—the deadliest in recorded history. For scientists and engineers, data from the tsunami provides a unique and irreplaceable data collection, invaluable for validating and improving the accuracy of models of tsunami behavior. More about the development and management of the Indian Ocean Tsunami Data Repository can be found on page 6.

Of course, the best time to understand the implications of a disaster is before it happens. In the last year, a consortium of geoscientists and earthquake engineers working with the Southern California Earthquake Center (SCEC) partnered with two dozen SDSC staff to run the largest scale simulation of a large magnitude earthquake in southern California to date. The "TeraShake" simulation modeled a 600 by 300 kilometer area (which included the Los Angeles basin) at 200 meter resolution (grid points every 200 meters). The original simulation ran for a week on SDSC's DataStar and produced 47 TeraBytes of data (over 4.5 times the printed materials in the Library of Congress). The simulations are ongoing at SDSC, producing hundreds of terabytes of data that can help lead to improved hazard estimates and safer structures.

All of us at SDSC would like to dedicate this issue to the scientists, engineers, and technologists in our community, both at SDSC and elsewhere at many institutions and agencies, who have worked so hard to lend their time and expertise during recent disasters. Your efforts have made an immense difference.



Hurricane Katrina, which devastated New Orleans and the Gulf coast in the costliest natural disaster in the nation's history, seen from a NOAA weather satellite as the eye makes landfall. SDSC database experts provided vital assistance to the American Red Cross by consolidating multiple lists of Katrina evacuees into a single, easy-to-use missing and found "superlist." NOAA/NESDIS.

Building Katrina's Superlist

Photos © American Red Cross

by Greg Lund

Reuniting Hurricane Evacuees
Hundreds of thousands of people were displaced and separated by Hurricane Katrina. SDSC data expertise helped reconnect them by providing unified missing and found lists to the Red Cross.

Hurricane Katrina crashed ashore on August 28th to become one of the most damaging storms in US history. Where the winds blew down homes, families scattered. Where floods inundated homes, people ran or swam for their lives. Thousands of people simply fled. They ended up in shelters, athletic arenas, or in homes across the Gulf states and across the country.

That flight started a great separation. Parents were separated from their children, and husbands separated from wives. Some were only a few miles from one another, but had no idea how to tell family and friends they were safe. Others ended up thousands of miles away from each other without a clue where to look. The Red Cross alone housed evacuees in 1,150 shelters across 27 states and the District of Columbia.

Efforts to reunite lost families began even before Katrina's winds and rain diminished. Missing persons lists sprang up all over, seemingly every place that housed evacuees. The Red Cross, newspapers, and websites all began independent lists to try to reunite loved ones. By the Red Cross' tally, more than 287,000 names would ultimately be registered online.

DATA EXPERTISE

SDSC couldn't do much to stem the floodwaters, but Dr. Chaitan Baru, a renowned data scientist at SDSC, and his team could do something about the deluge of missing persons' data threatening to choke available resources and dramatically slow down the reunification process.

It started with a phone call to Baru less than 24 hours after first reports of levee breaches in New Orleans. Dr. Eric Frost, an associate professor in the department of Geological Sciences at San Diego State University and an SDSC collaborator, was working with the National Institute for Urban Search and Rescue and the Red Cross to provide first responders before-and-after satellite images of devastated sections of the gulf coast. He called to ask for extra space in SDSC's large data storage facility for these images. Richard Moore, SDSC's director of production services, quickly agreed and marshaled staffers to provide server space for Frost's effort.

This was followed the next day by a request that would eventually leverage knowledge gained from research in

managing large data sets at SDSC. The Red Cross needed help in dealing with the growing data from missing person's lists. In the first few days after the hurricane dozens of separate lists were posted online, ranging from CNN, the International Committee of the Red Cross, and MSNBC, to smaller gulf coast newspapers. There was no uniformity in the information gathered, and one list might look completely different from the next but still contain many of the same names. So, the Red Cross asked for help in sorting 911 calls reporting missing people.

Baru enlisted the help of SDSC database experts Jerry Rowley and Vishu Nandigam to begin the task of amalgamating the names from more than 30 independent lists into one, coordinated and easy-to-use master list. A large task, much easier discussed than actually done.

That meant ditching any Labor Day weekend plans for the SDSC team, which had grown to include Neil Cotofana, Peter Shin, Hector Jasso, L.J. Ding, and Roger Unwin. Baru's team had significant experience in database management, but this task was unique. It started around a white board with the team trying to find all available missing persons lists, determine how to compile data from lists that differed substantially in the type of information contained, and then combine that data into one list. In addition, it was a race against time to get a list together as quickly as possible to start reconnecting families.

"We immediately deployed a data upload site where the Red Cross and other data providers could upload their files containing names of safe and missing people. We commandeered one of our computer systems at SDSC and used it to develop a database of all the names," said Baru. "At the same time, we requested some of our staff volunteers to begin investigating approaches for 'fuzzy' matching on names, since we knew that the incoming data was going to be of highly varying quality."

Jerry Rowley coordinated the daily activities of the group at SDSC. "Our original plan was to also provide a website to support searching the names in the single list that we were creating," said Rowley. "Over the Labor Day weekend, some of our staff were working on downloading, cleaning, and loading the data, while others developed a new website along with a search interface."

Above: SDSC data researcher Chaitan Baru led the Center's efforts to provide emergency data management help to the American Red Cross in responding to Hurricane Katrina. B. Tolo.

Left: SDSC database researcher Jerry Rowley coordinated day to day efforts of the SDSC team. B. Tolo.

The initial fuzzy search was based only on the first and last name of a person. To implement a more powerful search, SDSC contacted commercial vendors of fuzzy search software, and one of them, Identity Systems, offered the software for free use and immediately flew one of their customer support technicians from New York City to San Diego. SDSC researcher Doug Greer is using the software to efficiently match the names of missing people with the names in the amalgamated safe list at SDSC.

EFFECTIVE TEAMWORK

On Monday, September 5, as the refugee chaos was reaching its zenith, the collaboration between SDSC, the Red Cross, and eventually Microsoft kicked into high gear. The Red Cross had the largest list of names, Microsoft was developing a website, www.katrinasafe.com, and SDSC had the data expertise. Together the partners raced against time to access as much missing person data as possible, refine it, and place it on one readily-available and prominent website.

SDSC signed a nondisclosure agreement with the Red Cross, in effect promising to keep the data on individuals safe and sound. That paved the way for a deluge of Red Cross names gathered from the ever-growing number of shelters in the South.

"It was tremendously encouraging to see the spirit with which our staff responded to this call for help. People dropped their plans for the long weekend and simply went ahead with the tasks they were given with gusto. We pulled together a team across multiple departments at SDSC, and everyone worked together with great enthusiasm," said Baru. "We were also extremely impressed with the high degree of professionalism of the Red Cross staff, it was clear we were working with a group of individuals very committed to the cause of helping others."

The researchers were aware that the data they were dealing with was extremely sensitive. It contained information on individuals, including their addresses, phone numbers and, occasionally, other

private information. For example, in some cases individuals voluntarily provided their Social Security numbers. "As the central aggregator of all this data, we are extremely careful about how we treat this information," said Baru. "We 'sanitize' and 'anonymize' the data before passing it to others who are not from the Red Cross."

SDSC is filling a business-to-business (B2B) role in this effort, Baru explained. The amalgamated list of safe and missing names that SDSC creates is then provided to Microsoft, who is filling the business-to-customer role, providing access to this information to the public through the KatrinaSafe.com website.

"In our B2B role, SDSC is offering a search service to businesses. After we receive lists of names of missing people from private companies, government agencies, and nonprofits, we do a fuzzy match of these missing people names with our safe list, and then return that information directly to the requesting institution," said Baru. "This information is kept private between that institution and SDSC."

Whether thousands of miles apart or very nearby, separated families often had no way to find each other except Internet lists such as KatrinaSafe.com that SDSC staff helped create.

© American Red Cross

GOING PUBLIC

The partnership was a success. Two lists were developed, a list of the missing and a list of the found. Data from the amalgamated list started flowing from SDSC on Tuesday, September 6, and SDSC has since provided more than 486,000 records to the KatrinaSafe.com site, both safe and missing persons. As data continued to pour in, additional SDSC staff volunteered to help clean the data, including Geoff Avila, Linda Ferri, John Helly, Cynthia Lee, Jon Meyer, Hannes Neidner, Jane Park, Bob Sinkovits, and Donna Turner.

The search for evacuees took new forms. Large gulf coast area companies impacted by Katrina asked SDSC to mine the data lists to help locate their employees who had been scattered. The National Center for Missing and Exploited Children offered names of missing kids to be matched against the amalgamated data. Government agencies, too, took part in the massive effort to locate evacuees.

"SDSC provides a comprehensive set of data storage and analysis tools and technologies for the science and engineering research and education communities," said Dr. Francine Berman, SDSC Director. "All of the staff at SDSC want to help, and we are delighted that we can use our data tools and technologies to facilitate the difficult and important job of helping identify and reconnect Katrina's survivors."

Throughout the Katrina effort there was another thought in everyone's mind: Let's put together a data system to handle the next disaster. "This experience has clearly demonstrated the need for a rapid IT deployment capability during emergencies," said Baru. "We're glad that the capacity and expertise available at SDSC was able to serve the need in this case, and learning from this experience, we've decided to launch a new initiative at SDSC called SURGE, for SDSC URGent Response."

– Greg Lund is Director of Communications at SDSC.

URL: www.katrinasafe.com

SDSC SUPPORTS TSUNAMI DATA COLLECTION

Following the devastating Indian Ocean tsunami in December 2004, more than 20 National Science Foundation-funded scientific reconnaissance teams went to work in Asia, gathering voluminous data from the tsunami—the deadliest in recorded history. The data collected will be preserved, curated, and made available for use by researchers. This unique and irreplaceable resource will help them better understand the broad impact of tsunamis on communities, buildings, ecology, and people, as well as guiding preparations for future events.

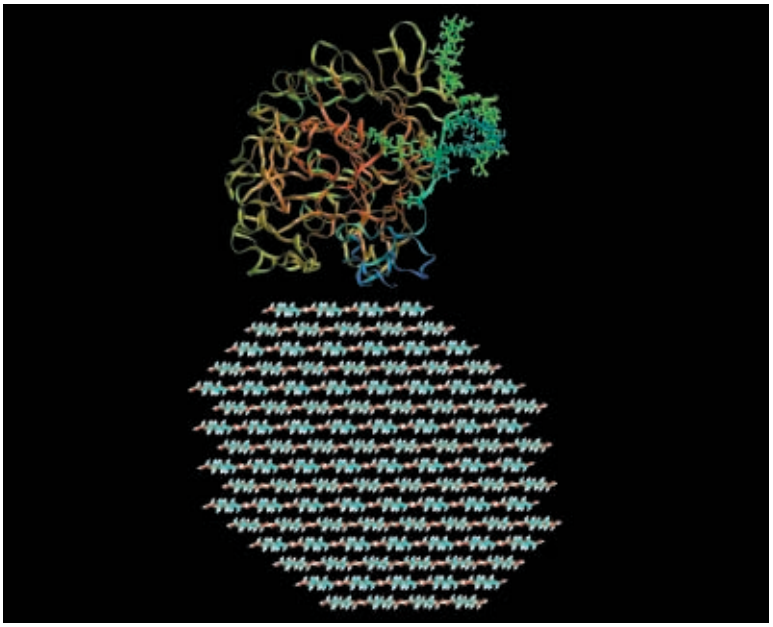
The data preservation effort is being led by SDSC's Network for Earthquake Engineering Simulation (NEES) Cyberinfrastructure Center (NEESit) team. This group operates and supports an extensive central information technology infrastructure for earthquake engineers and researchers.

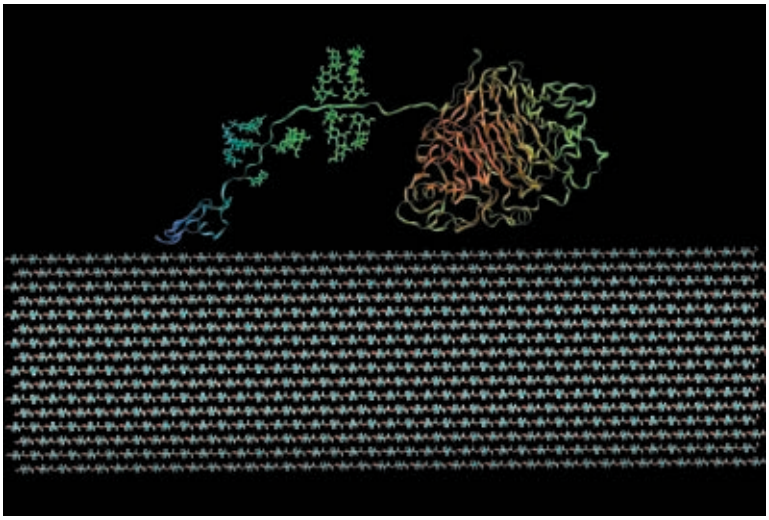
"The teams are working with regional partners to gather and translate data in support of research projects aimed at better understanding the impact of this unique event." said SDSC's Anke Kamrath, director of NEESit. "We're helping to capture and preserve these findings to support research endeavors on this tsunami which may continue for decades or even centuries from now." The tsunami data is scheduled to be made publicly available through the NEES Tsunami Data Repository at tsunami.nees.org starting in 2006.

The NSF-sponsored tsunami researchers have looked at many different types of tsunami impacts on the world. Teams focusing on collecting social, environmental, geological, ecological, and other kinds of data spent several months in the region. Information collected ranges from social data on human behavior and tsunami responses to remote sensing satellite data on ecological impacts and data on tsunami deposits and sedimentology.

Tapping Plants for Fuel

SDSC Helps Accelerate Cellulose Conversion to Ethanol





Cellulase Enzyme
Top and Side view of the enzyme cellulase, processing along the top of a fiber of the sugar polymer cellulose. The enzyme breaks cellulose into smaller sugar molecules called beta-glucose which are then fermented to make the renewable fuel, ethanol. J. Brady, M. Himmel, L. Zhong, M. Crowley, C. Brooks III, and M. Nimlos.

by Cassie Ferguson

Imagine mowing your lawn and then dumping the grass clippings into the gas tank of your car. Inside your tank, the grasses are digested and converted into ethanol—a high-performance, clean-burning, renewable fuel. You avoid the astronomical cost of filling up with old-fashioned petroleum, and the U.S. avoids the costly environmental, climate, and security issues of depending on nonrenewable fossil fuel. While tapping yard clippings as a source of gas might still be something found only in movies, the use of plant material as a major energy source has attracted nationwide attention, with ethanol blends already being offered at the pump.

But the process of producing ethanol remains slow and expensive, and researchers are trying to formulate more efficient, economical methods—a challenge that hinges on speeding up a key molecular reaction being investigated in a Strategic Applications Collaboration between researchers at SDSC, the Department of Energy's National Renewable Energy Laboratory (NREL), Cornell University, The Scripps Research Institute, and the Colorado School of Mines.

The interest in ethanol, commonly known as grain alcohol, is being driven by the combination of rising petroleum prices and government subsidies for so-called biofuel, a mixture of 15 percent (by volume) gasoline and 85 percent ethanol, known as E85, which sells for an average of 45 cents less per gallon than gasoline. Efforts to mitigate climate change are also spurring the growth of such renewable fuels, which add far less net greenhouse gas to the atmosphere than burning fossil fuels because the step of growing plant material removes carbon dioxide. In August 2005, President George W. Bush signed a comprehensive energy bill that included a requirement to increase the production of biofuels including ethanol and biodiesel from 4 to 7.5 billion gallons within the next 10 years.

While most people are familiar with the process used to turn plant material—such as hops—into ethanol-containing beverages like beer, that process is slow, expensive, and the end product too impure for energy use. To produce ethanol for energy use on a massive scale, researchers are trying to perfect the conversion of "biomass"—plant matter such as trees, grasses, byproducts from agricultural crops, and other biological material—via

PROJECT LEADERS

John Brady, Cornell
Mike Himmel, NREL

PARTICIPANTS

Mark Nimlos, Mike Himmel, NREL
James Matthews, John Brady, Cornell
Linghao Zhong, Penn State
Xianghong Qian, Colorado School of Mines
Michael Crowley, Charles Brooks, TSRI
Mike Cleary, Giri Chukkapalli, SDSC /UCSD

industrial conversion in "biorefineries."

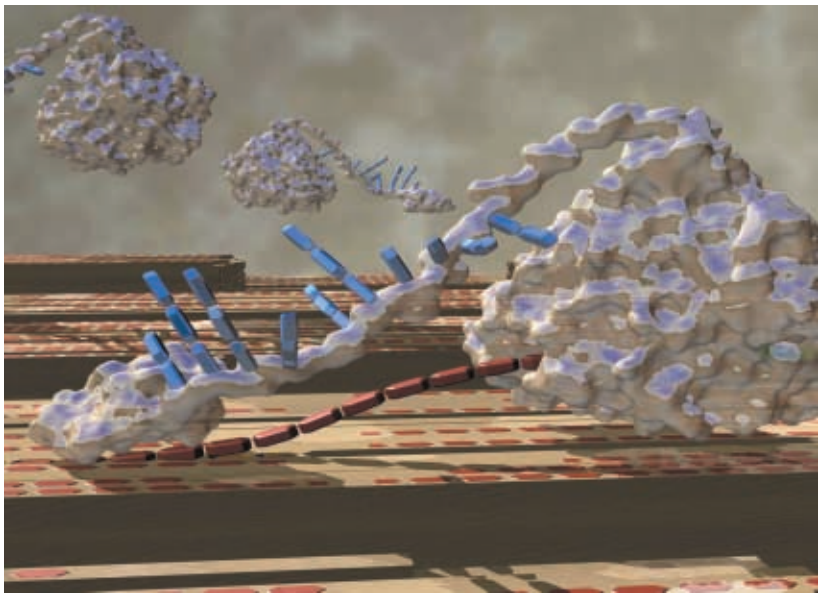
"Cellulose is the most abundant plant material on earth and a largely-untapped source of renewable energy," said project manager Mike Cleary, who is coordinating SDSC's role in the project. "So this collaboration is addressing not just a significant problem in enzymology but a problem of huge potential benefit to society."

MODELING A MOLECULAR MACHINE

The central bottleneck in making the biomass to fuel conversion process more efficient is the current slow rate of breakdown of the woody parts of plants—cellulose—by the enzyme cellulase, which also happens to be expensive to produce. The enzyme complex of cellulases, made up of proteins, acts as a catalyst to mediate and speed this chemical reaction, turning cellulose into sugars.

Scientists want to understand this process at the molecular level so that they can learn how to enhance the reaction. Using molecular dynamics simulations, which model the movement of the enzyme at the atomic scale, the researchers want to determine if the kinetics of the enzyme agree with models based on biochemical and genetic studies. By probing in minute detail how the enzyme makes contact with cellulose at the molecular level, the researchers hope to speed up the process and make it more cost effective by discovering ways the enzyme can be altered through genetic engineering.

The cellulase enzyme complex is actually a collection of protein enzymes, each of which plays a specific role in breaking down cellulose into smaller molecules of sugar called beta-glucose. The smaller sugar molecules are then fermented with microbes, typically yeast, to make the fuel, ethanol. One of the parts of the enzyme complex, cellobiohydrolase (CBH I), acts as a "molecular machine" that attaches to bundles of cellulose, pulls up a single strand of the sugar, and puts it onto a molecular conveyor belt where it is chopped into the smaller pieces. In order to make this process more efficient through bioengineering, researchers will need a detailed molecular-level understanding of how the cellulase enzyme functions. But the system has been difficult to study because it is too small to be directly observed under a microscope while too large for traditional molecular mechanics modeling.



Molecular Machine
In this artist's rendering, the large translucent enzyme cellulase, resembling a dinosaur, pulls up strands of cellulose, digests them in its belly, and excretes them as smaller pieces of sugar. M. Himmel, NREL, DOE Biomass Program.

To explore the intricate molecular dynamics of cellulase, researchers at NREL have turned to CHARMM (Chemistry at HARvard Molecular Mechanics), a suite of modeling software for macromolecular simulations, including energy minimization, molecular dynamics, and Monte Carlo simulations. The widely-used community code, originally developed in 1983 in the laboratory of Martin Karplus at Harvard University, models how atoms interact.

In the cellulase modeling, CHARMM is used to explore the ensemble configurations and protein structure, the interactions of the protein with the cellulose substrate, and the interactions of water with both. Not only are the NREL simulations the first to simultaneously model the cellulase enzyme, cellulose substrate, and surrounding water, they are among the largest molecular systems ever modeled. In particular, the researchers are interested in how cellulase aligns and attaches itself to cellulose, how the separate parts of cellulase—called protein domains—work with one another, and the effect of water on the overall system. And they are also investigating which of the over 500 amino acids that make up the cellulase protein are central to the overall workings of the "machine" as it chews up cellulose.

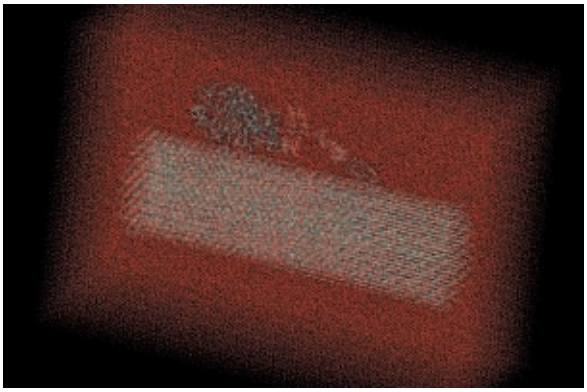
To the biochemists in the collaboration, the simulation is like a stop-motion film of a baseball pitcher throwing a curveball. In real-life the process occurs too quickly to evaluate visually, but by breaking down the throw into a step-by-step process, observers can find out the precise role of velocity, trajectory, movement, and arm angle. In simulations on SDSC's DataStar supercomputer, the researchers have modeled a portion of the enzyme, the type 1 cellulose binding domain, on a surface of crystalline cellulose in a box of water. The modeling revealed how the amino acids of the domain orient themselves when they interact with crystalline cellulose as well as how the interaction disrupts the layer of water molecules that lie on top of the cellulose, providing a detailed glimpse of this intricate molecular dance.

PUSHING THE ENVELOPE

The NREL cellulose model includes over 800,000 atoms, including the surrounding water, the cellulose, and the enzyme—an enormous structure to model computationally. According to the researchers, an accurate understanding of what is happening will require the capability to scale up their simulation to run for 50 nanoseconds in the reaction—an extremely long amount of time in molecular terms and highly demanding in computational terms (there are one billion nanoseconds in one second). To reach 50 nanoseconds, the researchers must calculate 25 million time-steps at two femtoseconds per time step (one femtosecond is one quadrillionth of a second).

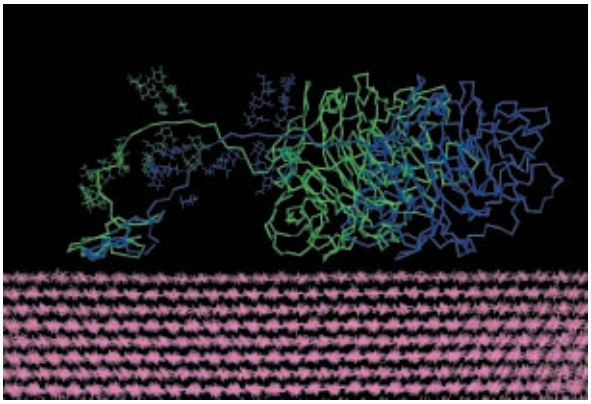
However, the sheer size of the model is beyond the limit of the current capabilities of the CHARMM simulation code, which has been difficult to scale as the number of computer processors grows larger, since the code was originally written to model thousands, not hundreds of thousands, of atoms. The SAC partners have worked to enhance CHARMM to scale to larger numbers of atoms and to run on some of the largest resources available to academic scientists in the U.S., including DataStar (recently expanded to 15.6 teraflops), TeraGrid (4.4 teraflops), and BlueGene (5.7 teraflops).

To determine how much time the large-scale CHARMM simulations require, a calculation on DataStar found that a series of 500-step simulations on a 711,887 atom system for one picosecond (one thousandth of a nanosecond) required 12 minutes on 64 processors and 9 minutes on 128 processors. Because of scaling issues, a full nanosecond run will require 1,000 times more time than these benchmarking runs, so that full-scale simulations are expected to require nearly one million CPU hours.



Top: CHARMM
Chemistry at HARvard Molecular Mechanics, or CHARMM, a suite of modeling software for macromolecular simulations, is used to model the complete system of the enzyme cellulase digesting the sugar cellulose in water (reddish haze). J. Brady, M. Himmel, L. Zhong, M. Crowley, C. Brooks III, and M. Nimlos.

Bottom: Cellulase Enzyme
Side view of the enzyme cellulase, processing along the top of a fiber of the sugar cellulose. J. Brady, M. Himmel, L. Zhong, M. Crowley, C. Brooks III, and M. Nimlos.



To extend the capabilities of the CHARMM simulation code to this unprecedented scale, SDSC's Giri Chukkapalli, a computational scientist, along with Scripps' Michael Crowley, a software developer in Charles Brooks' lab, have reengineered parts of CHARMM to be more efficient running as a parallel, rather than serial, application. In particular, the researchers in the SAC collaboration have targeted a number of subroutines in the code, which are being altered to speed up its performance on 256 and 512 processors.

A MODEL PARTNERSHIP

Outreach on the part of SDSC resulted in this large cross-agency collaboration based on a team approach, with interdisciplinary participation by biochemists from NREL, enzymologists and carbohydrate chemists from Cornell, software developers from TSRI, and computational scientists at SDSC. To validate and gauge the accuracy of the CHARMM simulations, the models are studied by James Matthews and John Brady of Cornell, and Linghao Zhong at Penn State, who compare the simulated version of the overall action of the cellulase complex with experimental results. Similarly, the chemists at NREL, including Mark Nimlos, Mike Himmel, and Xianghong Qian at the Colorado School of Mines, interpret the biochemical findings. In addition to assisting with the software development and scaling to be able to run larger simulations, SDSC is also the key site for computation since the center houses compute resources such as DataStar, with capabilities far beyond those available at the other collaborators.

"We were looking for opportunities for collaboration with other agencies," said Cleary. "SDSC has unique expertise to offer in improving community codes like CHARMM and other molecular dynamics tools like AMBER." It turned out that Cleary, along with other SDSC staff, knew some of the researchers who had been working on the cellulase problem at NREL and the other sites. Their work was an ideal fit for a SDSC SAC collaboration, with each group lending its expertise to the project.

The collaboration, funded by U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy, and the Office of the Biomass Program, fit the mission of SDSC's SAC program—to enhance the effectiveness of computational science and engineering research conducted by nationwide academic users. The goal of these collaborations is to develop a synergy between the academic researchers and SDSC staff that accelerates the researchers' efforts by using SDSC resources most effectively and enabling new science on relatively short timescales of three to 12 months. And beyond the project results, the hope is to discover and develop general solutions that will benefit not only the selected researchers but also their entire academic communities and other high-performance computing users.

In this case, beyond being able to model cellulase digesting cellulose to improve the production of ethanol, the improvements to CHARMM are opening the door so that the software, running on cutting-edge hardware systems, can simulate many other large-scale biological systems. In turn, that will allow scientists to pose entirely new questions, opening novel avenues for research, said Cleary.

According to Chukkapalli, "We're excited about the advent of new architectures that provide massive amounts of computing power. The questions from biophysics, structural biology, and biochemistry that have been only dreams in the minds of computational chemists are now on the verge of being studied in realistic simulations." *–Cassie Ferguson is a freelance science writer.*

RELATED LINKS

URL: National Renewable Energy Laboratory's Biomass Research Page
www.nrel.gov/biomass

SDSC Strategic Applications Collaborations Program
www.sdsc.edu/user_services/sac

REFERENCES

Mark R. Nimlos, Stan Bower, Michael E. Himmel, Michael F. Crowley, Giridhar Chukkapalli, Michael J. Cleary, and John Brady. (May, 2005) Computational Modeling of the Interaction of the Binding Domain of T. reesei Cel7A with Cellulose. Poster presentation at the 27th Symposium on Biotechnology for Fuels and Chemicals, Denver, CO.

Sheehan, J. & Himmel, M. (1999). Enzymes, energy, and the environment: A strategic perspective on the US Department of Energy's Research and Development Activities for Bioethanol. *Biotechnology Progress,* 15, 817-827.

swami
the Next Generation Biology Workbench

PARTICIPANTS

Mike Cleary, Kevin Fowler, Mark Miller, Gregory Quinn, Ashton Taylor, and Roger Unwin, SDSC/UCSD
Shankar Subramaniam, UCSD and SDSC/UCSD
Celeste Brown, U. Idaho

by Lynne Friedmann

A Tool for Education
Using the Biology Workbench in the Bioinformatics Teaching Lab at the University of Idaho. Michael Placke.

A free, Web-based interface that links major molecular biology databases with analysis programs—accessed thousands of times a week by scientists and students worldwide—is about to become even better.

In May, SDSC announced the award of $2.2 million from the National Institutes of Health (NIH) to build on the groundbreaking "Biology Workbench," introduced nearly a decade ago by researcher Shankar Subramaniam to provide broad access to many biology software tools and data resources through a Web-based, point-and-click computational environment. The Biology Workbench also offers the advantage of speed, letting researchers complete within hours work that once took days, weeks, or even months.

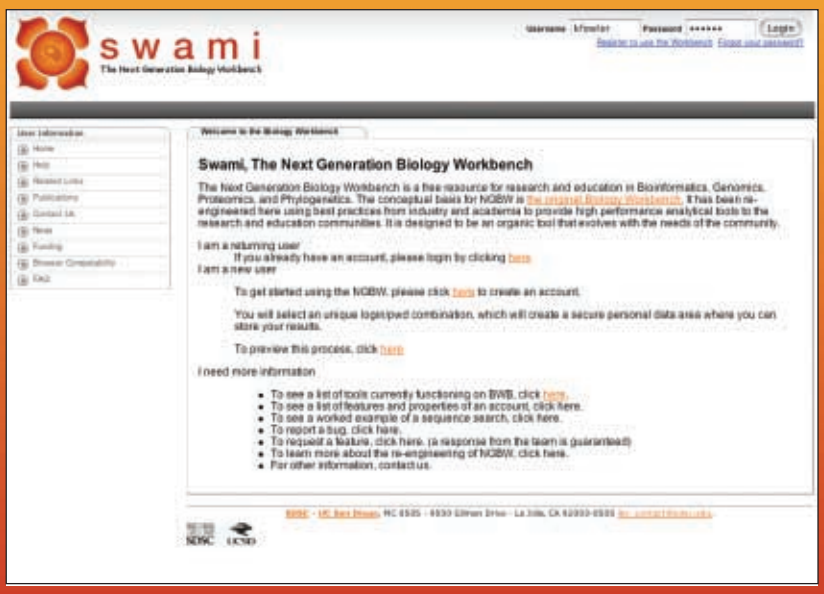
The Next Generation Biology Workbench (NGBW), also known as Swami, is a unique and versatile biological analysis environment that allows users to search up-to-date protein and nucleic acid sequence databases. Searching is integrated with access to a wide variety of analysis and modeling tools, all within a point-and-click interface that eliminates file format compatibility problems. This has made the Biology Workbench an indispensable resource for scientists who work in bioinformatics, an emerging discipline in which researchers collect, integrate, model, and analyze the explosion of biological data produced by such efforts as the Human Genome Project.

In addition to adding to fundamental scientific knowledge, this research can lead to improved understanding of disease and open the door to development of new treatments and drug discovery. The NGBW prototype, as well as links to the current Biology Workbench can be found at www.ngbw.org.

"I have been using Biology Workbench on a regular basis for the last three to four years," said Vanderbilt University assistant professor, Mark de Caestecker. "It has proved to be an invaluable tool for the analysis and design of gene and protein constructs used in a range of different experiments in my laboratory. The Biology Workbench has the most comprehensive and easy-to-use applications I have come across." In his research, de Caestecker studies stem cell differentiation in kidney development, cancer, and tissue injury repair, and also researches cellular signaling in relation to hypertension.

UPDATING A FAVORITE

The researchers are making the Next Generation Biology Workbench even more



Updating a Favorite: Next Generation Biology Workbench Interface.

useful by expanding its present offering of 65 tools. Like the original, the NGBW will continue as a free Web resource that offers access to data, data storage, software tools, and computational resources that help researchers mine the information in many popular protein and nucleic acid sequence databases. NIH funding will support the construction of up-to-date features such as improved user interfaces and an expandable architecture that will allow the NGBW to continue to evolve in the future in response to new developments in technology, biology, and the needs of scientists.

"There have been huge leaps in the technologies used in building cyberinfrastructure since the original Biology Workbench was created," said Mark A. Miller, SDSC project leader for the new grant. Work will be done in phases with a beta release planned for April 2006. According to Miller, there are some upgrades that can be accomplished in a matter of months, while others will take a year or two to accomplish. For example, the current workbench integrates information from 33 public databases, which are downloaded into a flat file. Using the less powerful technology of the flat file format places significant limitations on search functionality. Therefore, a major goal of the Next Generation project is to adopt a relational database format in which the information is broken down into tables and categories, which then allows more complex queries, or scientific questions, to be answered.

"Software developers always want to make things very elegant so they can later expand and make them more modular," said Miller. "We do want that, but we don't want to make people wait five years for the next product. So our focus is giving users something today and then making it more elegant underneath."

Other improvements will include enhanced visualization and data management capabilities, and to make sure that these services are available even to users with only a lower speed dial-up modem. This will enable a wide range of users to pose sophisticated questions, even if they don't have access to advanced computing resources.

While not losing sight of the researcher with limited financial resources for whom the Biology Workbench was originally developed, enhancements to the Next Generation Biology Workbench are expected to also capture more high-end users. The current workbench, running on a Sun computer, allows access to more than 32,000 active users, the majority of whom look at single sequences, submitting more than 120,000 requests for analysis, or jobs, monthly. This is not high demand, and consequently not very expensive computationally.

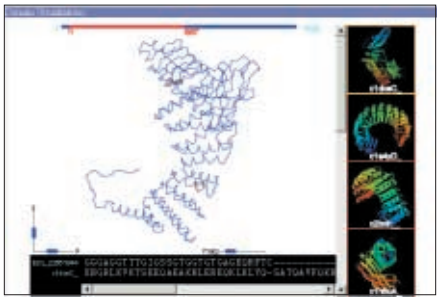
"We could probably handle four times that amount without breathing hard," said Miller. "But the current system is limited because of how the file system is structured. I believe we can design the Next Generation Biology Workbench so it will be fully expandable in the future."

A TOOL FOR EDUCATION

Because navigating the interface is comparable to learning the Windows or Macintosh operating systems, it didn't take long for instructors to embrace the original Biology Workbench as a teaching tool. Responding to this growing user segment, SDSC researchers are partnering with colleagues at the National Center for Supercomputing Applications (NCSA), where the workbench was initially developed, in developing an educational component. And consulting with educators in the early stages of the Next Generation Biology Workbench design, Miller explains, "will keep us from getting too far off the beam making a nice architecture but the wrong functions." The outcome will be a dedicated component or view for students and teachers, called the Student Biology Workbench.

This is welcome news to instructors like Celeste Brown, Bioinformatics Coordinator, Initiative for Bioinformatics and Evolutionary Studies at the University of Idaho. "Ten years ago a graduate student found the Biology Workbench (on the Web) and brought it to my attention," she said. "I've been using it for teaching ever since."

Brown is not alone. A Web search reveals a wide range of lesson plans designed specifically with the Biology Workbench in mind. Many are from smaller institutions of higher education, such as the University of Idaho. The use of the Biology Workbench in this setting should be welcome news to the NIH, which supports an initiative to encourage bioinformatics training in states that have historically not received as high a level of government grant dollars as, say, the state of California. In this program, 23 states and Puerto Rico qualify for additional NIH support for faculty development and enhancement of research infrastructure under the Institutional Development Award



Easy Visualization
The NGBW will feature lightweight server-side tools for visualizing and manipulating protein molecules, such as the scalable vector graphics viewer shown here. NGWB.

Facilitating Science
An example of the kind of protein rendering that users can easily produce with the Workbench. To make the Next Generation Biology Workbench usable by the most researchers and students, the project is developing tools that run on machines at SDSC so that users don't have to download and install software on their machines. NGBW.

(IDeA) Program. While there isn't a requirement that IDeA states use the Biology Workbench to train students in bioinformatics, many do.

At the University of Idaho, an evolutionary perspective is the emphasis of all biology training. "But let's face it. Nobody really likes development lab where you put an egg into a Petri dish and watch how it develops into a chicken," notes Brown. "A tool like the Biology Workbench puts things in context and reinforces scientific principles learned in earlier classes. Besides, students are used to a lot more technology than they were ten years ago."

Brown uses the Biology Workbench in an introductory enzyme lab course to access 3-D structures from the database. Students not only see what's going on in the reactions they set up, they can then consider how the structures evolved. "I want students to understand that there are databases out there that have all this nucleotide and protein sequence information," said Brown. "I also want them to realize that it's easy to get to and there are a lot of tools out there to help them analyze what's in those databases."

Outside the classroom researchers have become aware of the Biology Workbench through scientific publications. In many cases, at the end of a commercial software review, the Biology Workbench is mentioned as a free solution that scientists might also wish to consider. When the Next Generation Biology Workbench is ready for release, there is travel support in the NIH grant for a "roll out" of its new capabilities at a series of major national scientific meetings.

"The original Biology Workbench created a new paradigm for integrating biological information and tools, giving easy access to researchers as well as students," said Shankar Subramaniam, Professor of Bioengineering and Director of the Bioinformatics Program at UCSD. "It's rewarding to see the growing interest in this resource, and bringing the workbench forward using modern technologies will make this important tool more versatile and available to an even broader range of users."

When Miller recently posted an Internet request for testimonials about the Biology Workbench he soon heard from students, teachers, and researchers from the four corners of the earth. Feedback includes such superlatives as "invaluable," "easy to use," "faster and more efficient than other tools," and "critical for completion of my thesis research."

Professor of Biology Darrel Stafford at the University of North Carolina, Chapel Hill wrote, "We used it extensively in our search for the gene for *epoxide reductase* which was published in the March 5th issue of *Nature.*"

As far as Gabriel M. Belfort, an MD/PhD candidate at Boston University School of Medicine, is concerned, "not having the Biology Workbench would be the functional equivalent of replacing my computer with an abacus."

The Next Generation Biology Workbench team at SDSC includes Mark Miller, PhD, PI; Mike Cleary, PhD, co-PI and user advocate; Shankar Subramaniam, PhD, co-PI; Kevin Fowler, senior software architect; Roger Unwin, database engineer; Gregory Quinn, PhD, senior interface engineer; and Ashton Taylor, artist. Celeste Brown, PhD, of the University of Idaho, is education advisor and bioinformatics coordinator.

–Lynne Friedmann is a freelance science writer living in Solana Beach, California.

URL: www.ngbw.org

UNDER THE HOOD

A core goal of the project is to improve the Workbench using new technologies. The design goals include using an architecture that allows the NGBW to be freely available for distribution, and developed within the available budget and a short time frame. The team has addressed these issues by leveraging the Java Enterprise Edition software stack as implemented by the open-source JBoss 4.0 Application Server. The new workbench stores and retrieves data from relational databases by mapping Java objects to relational entities using JBoss' Hibernate persistence library. The user works with this data through a user-friendly, Web front-end that is implemented using the Apache Struts web-application framework.

"Because many users do not have authorization or sometimes the ability to install programs on their computers, we're developing very lightweight visualization tools that run effectively from the server side at SDSC," said Miller. This ensures that a wider range of users can benefit from NGBW data and tools.

Another challenge is how to support the wide variety of analytical tools made available within the Workbench, since such tools typically have very specific input and output format requirements. To do this, Miller explains, the developers must "wrap" each separate program, putting a translator between it and the central NGBW architecture, so that data supplied by the user can be passed to any of the analytical programs in the Workbench in the language it can interpret. In turn, the translator returns output to the user in a common format. And since the developers are using a well-defined common language, this also makes it straightforward for other developers to create their own tools to be added to or work with the Workbench.

REFERENCE
Subramaniam, S. The Biology Workbench—a seamless database and analysis environment for the biologist. *Proteins*, 32(1):1-2, 1998.

THE ELUSIVE NEUTRINO: New Window on the Violent Universe

SDSC's TeraGrid Data and Computing Resources Help Validate AMANDA Neutrino Telescope

Scientists have long sought ways to map the Universe and explore its most violent phenomena, from mysterious gamma ray bursts and supernovae to the black holes that inhabit active nuclei in the centers of galaxies. In their quest, some researchers are now focusing on subatomic particles called neutrinos, which show promise of being valuable messengers. In contrast to other particles or light, which are absorbed, bent, or scattered in their travels, the tiny, almost massless neutrino is able to travel virtually unimpeded across the vast distances of space to reach the Earth. Aided by the TeraGrid network, cluster, and massive data resources at SDSC at UC San Diego, physicists are developing a new kind of telescope, AMANDA-II, the Antarctic Muon and Neutrino Detector Array, to observe these neutrinos and decipher their tales about the location and inner workings of the cataclysmic events in which they originated.

Researcher Andrea Silvestri and Professor Steven Barwick of the Physics and Astronomy Department at the University of California, Irvine (UCI), along with many other scientists, are beginning to use the multipurpose AMANDA-II high-energy neutrino telescope at the South Pole to seek answers to a broad array of questions in physics and astrophysics. Observing neutrinos can shed light on fundamental problems such as the origins of cosmic rays, the search for dark matter and other exotic particles, as well as serving as a monitor for supernovas in the Milky Way.

However, the same qualities that let neutrinos travel freely across the universe also make them extremely difficult to detect. To further complicate matters, the vast majority of neutrinos that reach the Earth are produced nearby through cosmic ray collisions in the Earth's atmosphere, potentially masking the rarer, distant-origin neutrinos the scientists are seeking. How can particle astrophysicists Silvestri and Barwick filter out the mass of unwanted information from their data and tease out the tiny signal of high-energy neutrinos from far away?

by Paul Tooby

TAMING A FLOOD OF DATA

Scientists have steadily increased the size and effectiveness of the AMANDA telescope since it began collecting data in 1997. In AMANDA-II, the data acquisition electronics were upgraded with Transient Waveform Recorders that capture the complete waveform for each event detected. The researchers expect that several important goals will benefit by as much as a factor of 10 from the additional information gathered, including improvements in reconstructing muon cascades, the search for diffuse sources of ultra-high energy neutrinos, and the search for neutrino point sources.

However, with this progress in gathering data come new challenges, and the telescope now produces a flood of information, growing from one terabyte to 15 terabytes per year even in compressed form—about the same amount of information as in the entire printed collection of the Library of Congress or the data on 3,500 DVDs.

To analyze this immense data collection, Silvestri and Barwick turned to the large-scale data and computing capabilities of the NSF TeraGrid facility at SDSC. "The more capable the AMANDA telescope becomes, the more information we gather about neutrinos, which greatly helps our science," said Silvestri. "But this also means that to analyze all this data, we need the expertise and high-end resources of SDSC and the TeraGrid."

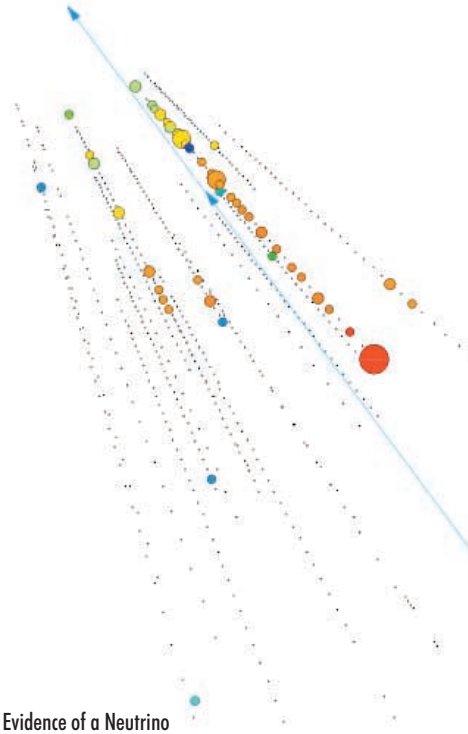
The first step was to transfer the 15 terabytes of raw AMANDA neutrino data over a high-speed network from UCI into a Storage Resource Broker (SRB) data archive at SDSC. Having developed the advanced SDSC SRB data management tool and installed more than one petabyte of online disk and more than six petabytes of archival storage capacity, SDSC is ideally suited to house and analyze massive data sets. And as the TeraGrid was designed to do, the high-speed network allowed the researchers to transparently access their massive data archive housed at SDSC for use on TeraGrid computational resources at various sites, speeding their research.

PROJECT LEADERS

Steven Barwick and Andrea Silvestri, UCI

AMANDA II Neutrino Telescope

To probe the most violent events in the Universe, scientists use the NSF-funded AMANDA neutrino telescope to seek evidence of elusive high-energy neutrinos. Aided by SDSC and TeraGrid data and compute resources, they successfully analyzed 15 terabytes of data. AMANDA, UW-Madison, photo Robert Morse.



Evidence of a Neutrino
One of the 1,112 observed high-energy neutrinos detected over one year as it traveled through the AMANDA instrument. In this side view of the detector looking upward, the lines of small black dots represent the strings of photo sensors, with colored circles marking sensors that detected light in this event (larger circles indicate more intense light). The earliest-arriving pulses (red light) are lowest in the array with later-arriving (green) light detected higher up, indicating an upward-travelling neutrino that passed through the Earth to reach the detector. A. Silvestri, UCI.

FINDING A NEUTRINO IN A HAYSTACK

The 15 terabytes of the full AMANDA-II waveform data collected for one year during 2003 contains some two billion experimental events, and the challenge the scientists faced was to identify the few neutrinos among the millions of times greater number of background muon events. Scientists measure neutrinos by detecting muons, which are subatomic particles produced in the rare interaction of a neutrino with other matter.

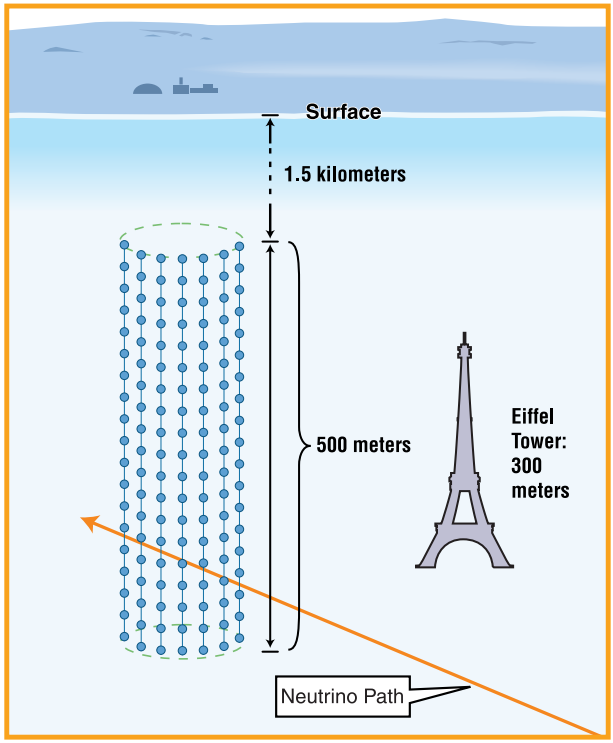
In their analysis, the researchers processed and filtered the experimental data and reconstructed each individual event. By running sophisticated algorithms on the TeraGrid through numerous iterations using likelihood-based statistical methods, the researchers analyzed the full 15 terabytes of experimental data. This process was highly data and compute-intensive, and only by having access to the resources of the massive SDSC online disk and tape storage and some 70,000 CPU hours on the TeraGrid supercomputer were the researchers able to carry out their data analysis. Typical jobs ran on 512 processors using one to two gigabytes of memory.

Finally, the researchers succeeded in distinguishing the faint signal of 1,112 atmospheric neutrinos from the billions of extraneous events. When they compared their results to standard analyses they found good agreement, confirming that the AMANDA instrument and SDSC-aided data analysis can produce the same physics results as previous data. In particular, the angular distribution of the atmospheric neutrino sample extracted from the standard data set agreed well with the new AMANDA data, with all 1,112 neutrinos originating in the northern hemisphere and distributed across the sky in a fairly uniform way, as expected.

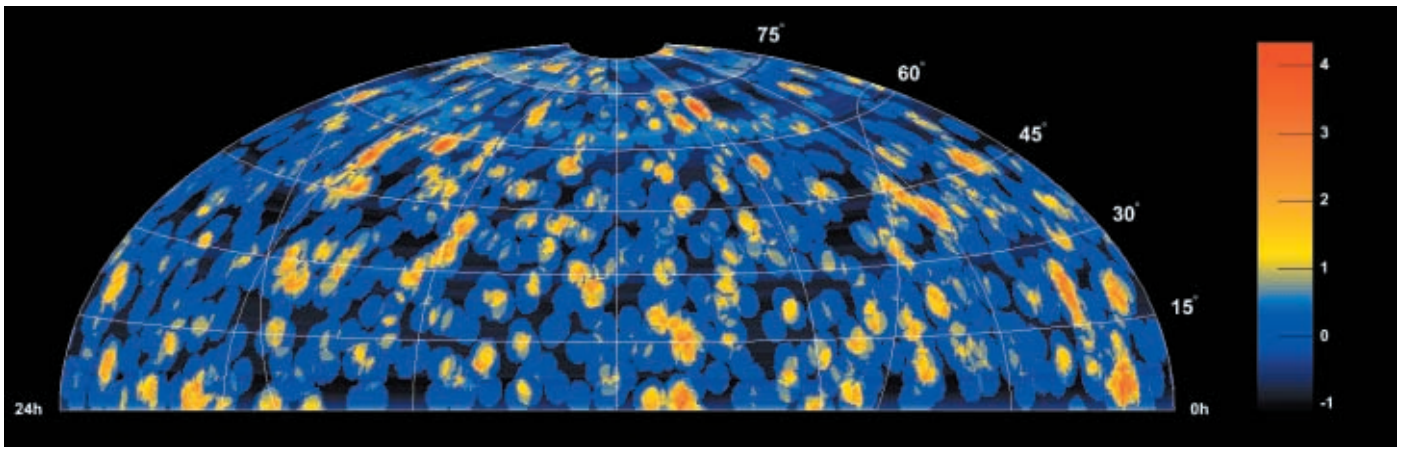
In validating the AMANDA instrument and analysis, the researchers also investigated whether the observed neutrinos were generated from collisions with the Earth's atmosphere, which produces a uniform spatial distribution of neutrino events across the sky, or whether some of the neutrinos were created by a source of extraterrestrial origin, which would be expected to produce a more concentrated event cluster in the direction of the source. Their statistical analysis of the data showed that the observed regions of the sky were compatible with atmospheric neutrino events, without significant event clusters that might indicate an extraterrestrial source.

NEUTRINOS: A NEW WINDOW ON THE UNIVERSE

Neutrinos are subatomic particles of neutral electrical charge and very small mass. Moving at nearly the speed of light, they participate only in weak and gravitational interactions, and can thus travel virtually unimpeded, giving a new view of the Universe. However, this also makes them very difficult to detect. When a high-energy muon-neutrino has a rare collision in the ice of the AMANDA detector, a muon particle is emitted that initially travels faster than the local speed of light in the ice. The muon radiates a sort of "shock wave" of light—analogous to the sonic boom of a supersonic aircraft. Traveling through the clear ice to the photodetectors of the AMANDA neutrino telescope, this Cherenkov light can reveal the presence of a high-energy neutrino. By correlating the light's arrival time in the 3-D photodetector array, scientists can pinpoint the direction of high neutrino-emitting objects, which can then be further studied.



Using Ice to Measure Neutrinos
The Antarctic Muon and Neutrino Detector Array (AMANDA), buried a mile deep in the Antarctic ice, is a new kind of telescope. Spanning a volume three times the size of the Eiffel Tower, it uses the clear ice without background light as a detector for sensitive measurements of rare neutrinos. Dan Brennan, UW-Madison.



Neutrino Sky Map

Sky map from the AMANDA neutrino telescope shows that the locally produced atmospheric neutrino background detected to date is quite uniform, without evidence of strong sources. This is part of the first validation of the new AMANDA-II neutrino telescope and the researchers' analysis system, which relied on large-scale TeraGrid data and compute resources at SDSC. Scale at right indicates excess from mean background events. A. Silvestri, UCI.

The scientists explained that it has been an enormous undertaking, requiring many years and the efforts of diverse groups and specialities working together to develop and validate an entirely new kind of telescope such as AMANDA. "This is a major result for the AMANDA-II neutrino telescope and broader research community," said Silvestri. "It's the first validation that we can in fact perform valid neutrino analysis with the new generation of instrument, its much larger data stream, and all the steps of our analysis, and we couldn't have done it without SDSC and the TeraGrid data and compute resources."

Moreover, since their initial analysis used only part of the complete information contained in the AMANDA-II waveform data, the researchers are now developing new software tools to exploit the full information available. The scientists expect the additional information to improve their ability to resolve even smaller differences in energy and angle. This will be crucial in their continuing search for the hard-to-detect energetic extra-terrestrial neutrinos that may hold the answers to many fundamental questions about the Universe.

A NEW KIND OF TELESCOPE

The fulfillment of a 40-year dream, AMANDA was designed to overcome the obstacles to detecting elusive neutrinos, and shows promise of giving scientists a startlingly broader view the Universe through the window of these high-energy particles. AMANDA is an ingenious new kind of "telescope" that senses neutrinos instead of light from above as have all telescopes since the time of Galileo (see sidebar). And unlike normal telescopes, which always face upward, AMANDA can also look downward, using the size of the Earth to "filter out" the extraneous downward-moving atmospheric muons, which are about a million times more abundant, and in this way detect high-energy neutrinos in the intermediate range from distant parts of the Universe. Only such neutrinos are able to pass through the whole Earth after entering in the Northern Hemisphere to reach the AMANDA telescope at the South Pole.

Occasionally, one of these upward-moving high-energy neutrinos will interact with an oxygen atom in the ice near the AMANDA array to produce a cascade of light-emitting muon particles. This light can travel long distances through the clear ice at the South Pole, which is free of competing background light, until it is picked up by the sensitive AMANDA photodetectors that gather this indirect evidence of the passage of a neutrino. The telescope can also search for even more energetic neutrinos by looking for downward-moving neutrinos of ultra-high energies.

The AMANDA neutrino telescope continues to grow in power, and currently consists of some 700 photon detectors arranged like beads on vertical strings, lowered into 19 holes in the ice at the South Pole. The holes are distributed across a circular area, creating a cylindrical volume of ice that serves as the detector some 120 meters in diameter and 500 meters tall, with its top about 1,500 meters below the ice cap's surface. Each photon detector module consists of a photomultiplier tube housed in a tough, pressure-resistant hollow sphere, with electrical and optical connectors attached. After a hole is bored in the ice with heated water, the string of detector modules is lowered into the water-filled hole, which then freezes solid, locking the detectors permanently in place.

In the future, the research will be scaled up even further in the NSF IceCube project, a much larger one-kilometer cube telescope array that will produce 20 times as much data, one terabyte per day or some 300 terabytes annually. Silvestri points out that "this will drive the need for even larger data, computational, and network resources at SDSC to better answer questions about the most energetic events in the history of the Universe."

– Paul Tooby is a senior science writer at SDSC and editor of EnVision Magazine.

REFERENCES

A. Silvestri et al, Performance of AMANDA-II using data from Transient Waveform Recorders, Proceedings of 29th International Cosmic Ray Conference, Pune, India, August 3-10, 2005.

A. Silvestri et al, The AMANDA Neutrino Telescope, Proceedings of International School of Cosmic Ray Astrophysics. Erice, Italy, July 2-13, 2004.

RELATED LINKS

Antarctic Muon and Neutrino Detector Array (AMANDA)
amanda.uci.edu

Andrea Silvestri
www.ps.uci.edu/~silvestri

Steven Barwick
www.ps.uci.edu/physics/barwick.html

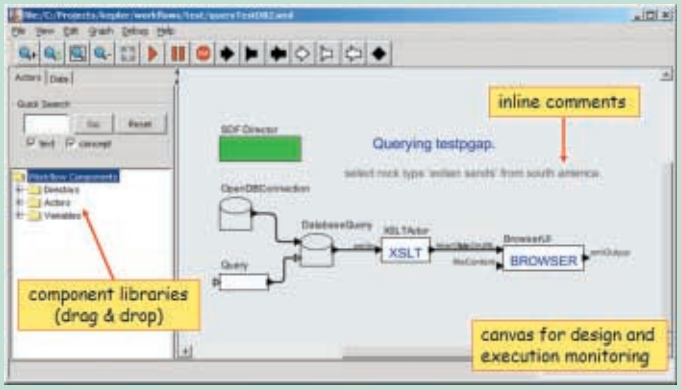
Faster Workflows:

Scientific Workflow Automation with Kepler

SDSC Helps Open Source Tool Streamline Scientists' Tasks and Aid Collaboration

The advent of the World Wide Web has brought a whole new world of data, tools, and online services within the reach of scientists. But this wealth of opportunities also brings a new set of challenges, and researchers can be overwhelmed by the sheer volume and complexity of resources. To overcome this problem, researchers at SDSC at UC San Diego, the National Center for Ecological Analysis and Synthesis (NCEAS) at UC Santa Barbara, and their partners have initiated an interdisciplinary collaboration to develop Kepler, a tool for scientific workflow management. By helping organize and automate scientific tasks, Kepler lets scientists take full advantage of today's complex software and Web services.

"For scientists, a good workflow tool is invaluable," said Kim Baldridge, a computational chemist at SDSC and professor at the University of Zurich. "It relieves researchers of the drudgery of tedious manual steps, but more importantly it can dramatically expand our ability to think bigger and ask new questions that were simply too complex or time-consuming before." Her research group has developed modules known as "actors" to carry out workflows in Kepler, leading to published results in the RESURGENCE project, which makes use of computational grids and Web services in computational chemistry.



Kepler Interface

The visual programming interface in the Kepler scientific workflow tool makes it easy for scientists to create low-level "plumbing workflows" to move data around and start jobs on remote computers as well as high-level data analysis pipelines that chain together standard or custom algorithms from different scientific domains.

Kepler is named after the Ptolemy software from UC Berkeley on which it is built, and the new tool is part of the emerging cyberinfrastructure, or integrated technologies for doing today's science. The open-source Kepler project grew out of the need for an analytical workflow tool in the Science Environment for Ecological Knowledge, or SEEK, project. The researchers from SDSC and UCSB who initiated Kepler worked with partners in the Department of Energy Scientific Discovery through Advanced Computing Program. Kepler has now grown to include more than half a dozen scientific projects, and the researchers plan to release a beta version in the coming months.

In its simplest form, Kepler may be thought of as a sort of "scientific robot" that relieves researchers of repetitive tasks so that they can focus on their science. In addition to increasing the efficiency of scientists' own workflows, Kepler will also give researchers increased capabilities to communicate and work together—searching for, integrating, and sharing data and workflows in large-scale collaborative environments.

"With Kepler, scientists from many disciplines can automate complex workflows, without having to become expert programmers," said Bertram Ludäscher, one of the initiators of the Kepler project and an SDSC Fellow and associate professor of Computer Science at UC Davis. "Kepler's flexibility and its visual programming interface make it easy for scientists to create both low-level 'plumbing workflows' to move data around and start jobs on remote computers, as well as high-level data analysis pipelines that chain together standard or custom algorithms from different scientific domains. And beyond automation, being able to document and reproduce workflows is a major objective of scientific workflow systems like Kepler."

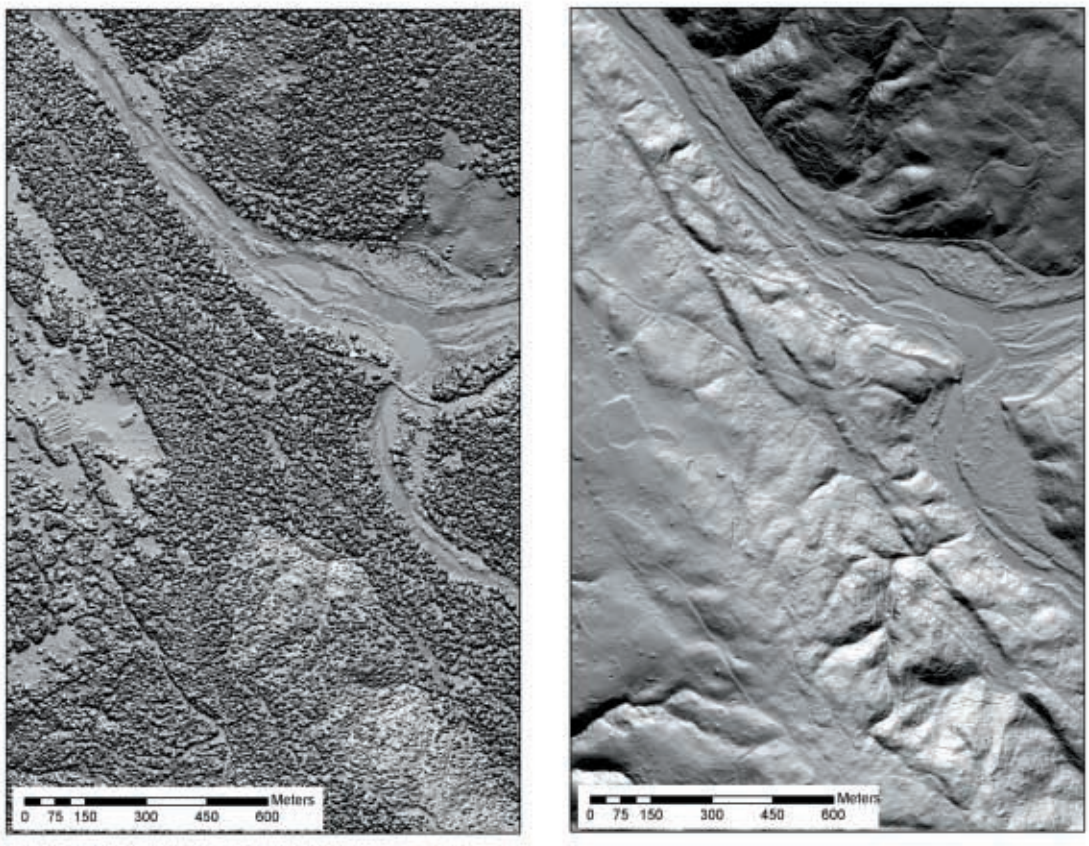
An important factor in ensuring that Kepler will be broadly useful across multiple scientific disciplines is its organization as an open source consortium. Participants in an open source project collaborate in building, maintaining, and peer-reviewing a common software tool. The source code is made publicly available without charge, and those who use the software are encouraged to contribute to its development—finding and fixing errors and adding new features that benefit the entire community. The Linux operating system and tools from the Apache Software foundation are well-known examples of open source efforts.

PROJECT PARTICIPANTS

Ilkay Altintas, Kim Baldridge, Zhijie Guan, Efrat Jaeger-Frank, Nandita Mangal, Steve Mock, SDSC/UCSD
Shawn Bowers, Bertram Ludäscher, UC Davis
Chad Berkley, Daniel Higgins, Matt Jones, Jing Tao, UCSB
Christopher Brooks, Edward A. Lee, Stephen Neuendorffer, Yang Zhao, UCB
Zhengang Cheng, Mladen Vouk, NCSU
Tobin Fricke, U. of Rochester
Timothy McPhillips, NDDP
A. Town Peterson, Rod Spears, U. Kansas
Kim Baldridge, Wibke Sudholt, U. Zurich
Terence Critchlow, Xiaowen Xin, LLNL

by Paul Tooby

Revealing the Earth Light Detection And Ranging (LIDAR) data is an important new tool for studying the Earth's surface, especially where heavy vegetation makes traditional aerial photography ineffective. Kepler workflows in the GEON project aid analysis of multi-terabyte LiDAR data. Image of the San Andreas fault in California shows tree canopy (left) from the first return signal and "virtual deforestation" (right) based on the last return signal, making the fault line clearly visible, running N.W. to S.E. C. Crosby, ASU.



"The fact that Kepler is open source encourages researchers to join the collaboration and build their own components, leveraging the infrastructure, and providing the vitality of a community approach to more rapidly extend Kepler's capabilities," said Ilkay Altintas, director of SDSC's Scientific Workflow Automation Technologies (SWAT) lab, which brings together scientific workflow efforts at SDSC under one umbrella. Researchers interested in scientific workflow technologies are invited to contact the lab to learn more.

TOWARD A GENERIC TOOL

When scientists search for relevant data sources and then undertake multi-step workflows, they must typically carry out and keep track of these complex steps in manual, ad hoc ways as they export and import data from one step to another across diverse environments. As a first step toward automating these tasks, scientists and computer scientists may collaborate on building a custom workflow tool. But this is an expensive and time-consuming process in which the software must generally be developed and maintained in an individual effort for each application.

To overcome these limitations, the Kepler initiative is developing a generic tool and environment that builds on existing technologies and will work in a wide range of applications to capture, automate, and manage researchers' actions as they carry out scientific workflows. The initial effort has brought together computer scientists with domain scientists in the disciplines of ecology, biology, chemistry, oceanography, geosciences, nuclear physics, and astronomy.

"With Kepler, we're giving scientists an intuitive tool that they can use to build their own workflows, which can include emerging Grid-based approaches to distributed computation," said Kepler co-initiator Matt Jones, a co-principal investigator and project manager for the SEEK project. "And in order to build a workflow environment that is effective across multiple domains of science, we're working with a growing range of projects to ensure the widest possible usefulness of the infrastructure."

In addition to the Ptolemy project of UC Berkeley described below, which serves as the framework for Kepler, the collaboration currently includes the following projects that span a range of scientific fields:

- Ecology: SEEK
- Biology, Nuclear Physics, and Astrophysics: Scientific Discovery through Advanced Computing Program in the Department of Energy (DOE SciDAC)
- Biology: NSF Encyclopedia of Life project (EOL) at SDSC
- Geosciences: the NSF GEON project, building a cyberinfrastructure for the geosciences
- Environmental science and sensor networks: the NSF Real-time Observatories, Applications, and Data management Network (ROADNet)
- Computational chemistry: the NSF RESURGENCE project
- Data mining tools: Cyberinfrastructure Laboratory for Ecological Observing Systems (CLEOS) at SDSC
- Distributed data integration: National Laboratory for Advanced Data Research (NLADR), a collaboration between SDSC and the National Center for Supercomputing Applications (NCSA).

Kepler is used in a wide variety of ways in these projects. In the Encyclopedia of Life project, the integrated Genome Annotation Pipeline software uses the Application Level Scheduling Parameter Sweep Template (APST) in month-long grid computing jobs that would be far more difficult without a workflow tool. First, Kepler prepares the databases and submits the computing job. Then it continues in a monitoring mode that checks on the execution and updates the corresponding database. In the event of a failure, the most recent update can be retrieved from the database, greatly

simplifying recovery. Kepler also makes it easy for scientists to execute a new task simply by double-clicking on and changing the parameters of an existing task. All of these capabilities let scientists accomplish genomic research much more rapidly.

To advance genomic research, biologists in SciDAC Program at the DOE study co-regulated genes. In their research, they try to identify promoters and develop models of transcription factor binding sites that play key roles in the expression of genes. The scientists use Kepler to help execute a series of data analysis and querying steps in which they move the results of each successive step from one Web resource to another. By automating these steps, the researchers save hours or days of time, speeding their results and allowing them to undertake problems on larger scales than previously possible.

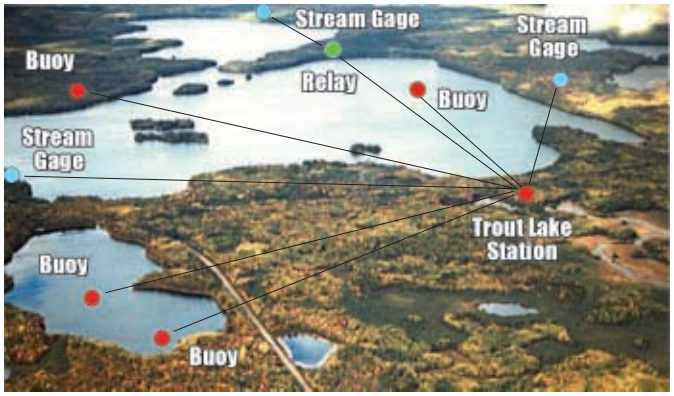
Ecologists in the SEEK project study problems that include invasive diseases such the West Nile virus. West Nile virus is spread through mosquitos feeding on migrating birds in a complex dual-vector process. The researchers develop predictions for where and how fast this kind of disease will spread. To do this, ecologists access online data sets about where the mosquitoes and birds are observed to live and migrate. Then they use Web-based ecological niche modeling tools to correlate this information with climate data, computing predictions for where the birds and mosquitos are likely to be found. Automating these steps with Kepler can make it feasible to produce accurate predictions for the spread of an invasive disease far more quickly than previously possible.

Automating workflows can yield similar benefits in a wide range of other scientific fields, and a growing number of projects and individuals are contributing to the Kepler open source project. More information on members and contributors can be found on the Kepler website at http://kepler-project.org/.

LEVERAGING EXISTING TECHNOLOGY

To explore whether they could build on existing technologies, the Kepler team surveyed available tools. Ptolemy, a project of the Center for Hybrid and Embedded Software Systems led by Professor Edward Lee of UC Berkeley, focuses on modeling, simulation, and design of concurrent, real-time embedded computing systems. The Kepler team realized that although Ptolemy had been developed for a very different purpose, it had capabilities that would provide a mature platform for the needs of Kepler in designing and executing scientific workflows. Ptolemy II, published as open source software, is the current base version of the Kepler infrastructure.

Ptolemy provides a set of Java language packages that support heterogeneous, concurrent modeling, design, and execution. Among Ptolemy's strengths are support for a number of precisely-defined models of computation such as streaming, and a concurrent dataflow paradigm for process networks that is appropriate for modeling and executing many scientific workflows. Ptolemy's programming approach is activity-based, or "actor-oriented" in Ptolemy terminology, which makes it easier to design the reusable components that scientists need. Ptolemy also has an intuitive graphical user interface called Vergil that allows users to compose complex workflows simply by stringing together individual actors, linking them according to the flow of data, and nesting them to represent desired levels of abstraction. In addition to Ptolemy's considerable built-in capabilities, which include more than 100 actors (or processing components) and directors (or workflow engines), the Kepler collaborators are continually adding new ones that extend the system, and have already contributed more than 100 additional actors.



Monitoring Lake Ecology
The SDSC Cyberinfrastructure Laboratory for Ecological Observing Systems (CLEOS) lab helps researchers gather data on lake ecology in the Collaborative Lake Metabolism Project. Kepler workflows can speed gathering, managing, and analyzing this complex multi-sensor streaming data, giving scientists new insights into lakes, including rapid events such as overturning. CLMP.

AUTOMATION CHALLENGES

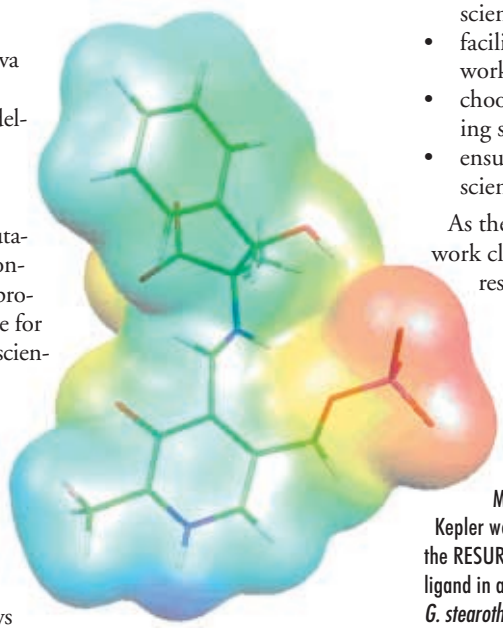
To capture the actions that scientists carry out in conducting their research and to automate these steps, the flow of data from one analytical step to another is described in Kepler in a formal, computer-readable workflow language. Among technical issues the researchers are facing in developing Kepler are:

- identifying factors that improve interoperability such as supporting shared workflow component repositories between Kepler, Taverna, Triana, SCIRun, DiscoveryNet, GeoVista Studio, and others
- managing heterogeneous mixtures of models of computation, including controlling space, time, and context
- handling distributed computation and code migration in scientific workflows
- facilitating the efficient use of existing scientific codes in workflow environments
- choosing or developing languages that are suitable for representing scientific workflows
- ensuring usability of the Kepler interface for the diverse group of scientists adopting Kepler.

As they resolve these technical challenges, Kepler developers must work closely with domain scientists in order to ensure that the resulting software meets the scientists' needs.

ENHANCING COLLABORATION

Beyond automating the steps of a given project, workflows captured in Kepler are intended to promote communication and collaboration for scientists in diverse



Modeling Molecules
Kepler workflows are being tested in studies involving this ligand/protein system in the RESURGENCE project. Image shows molecular electrostatic potential map of a ligand in a binding site of a mutated enzyme PLP-dependent alanine racemase from *G. stearothermophilus.* K. Baldridge and C. Amoreira.

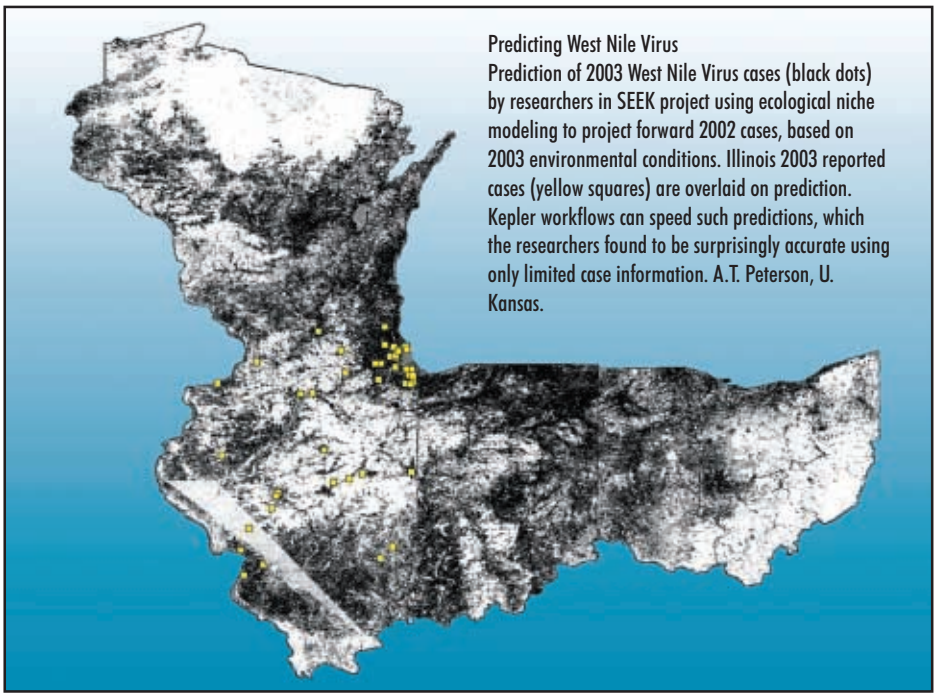
domains—a crucial capability for today's large-scale interdisciplinary collaborations. "Through its systematic approach to scientific workflows, Kepler can fulfill the important function of publishing analyses, models, data transformation programs, and derived data sets," said Kepler co-initiator Jones. "This gives scientists a way to track the provenance of derived data sets produced through workflow transformations, which is essential to being able to identify appropriate data sets for integration and further research."

In addition to distributing new Kepler actors that automate specific tasks, scientists can publish the results of workflows, storing the formal workflow descriptions of the steps carried out in a Web-accessible repository such as one of the metadata catalogs that are part of the SEEK EcoGrid. Kepler developers are working on extensions that will allow scientists to easily publish their workflows and share them with colleagues in flexible ways.

"We're now starting to add semantic capabilities to Kepler," said Shawn Bowers, project scientist at the UC Davis Genome Center where he works with professor Ludäscher on workflow and data integration technology for the SEEK project. "These include domain-specific ontologies acting as 'semantic types' for datasets, which will let scientists use the concepts of their own fields to search for and discover data and services, link to, and integrate data sets in both local and distributed grid environments."

Scientists are also interested in the potential of Kepler and related tools to power comprehensive "science environments," which they envision will follow accelerating growth paths as scientists are rapidly and seamlessly able to find out about and build on the previous work of their own and collaborating groups. "As the Kepler environment gains momentum and becomes more robust and reliable," explains SDSC's Altintas, "the body of resources that scientists can build upon grows larger, and more groups and scientific domains are joining this open collaboration."

– Paul Tooby is a senior science writer at SDSC and editor of EnVision Magazine.



Predicting West Nile Virus
Prediction of 2003 West Nile Virus cases (black dots) by researchers in SEEK project using ecological niche modeling to project forward 2002 cases, based on 2003 environmental conditions. Illinois 2003 reported cases (yellow squares) are overlaid on prediction. Kepler workflows can speed such predictions, which the researchers found to be surprisingly accurate using only limited case information. A.T. Peterson, U. Kansas.

RELATED LINK

URL: Kepler Scientific Workflow Project
kepler-project.org

REFERENCES

I. Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, S. Mock, Kepler: An Extensible System for Design and Execution of Scientific Workflows, 16th International Conference on Scientific and Statistical Database Management (SSDBM'04), 21-23 June 2004, Santorini Island, Greece.

B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger-Frank, M. Jones, E. Lee, J. Tao, Y. Zhao, Scientific Workflow Management and the Kepler System, *Concurrency and Computation: Practice & Experience,* Special Issue on Scientific Workflows, to appear.

West Nile Virus
In the United States, West Nile virus is transmitted by infected mosquitoes, primarily members of *Culex* species. © 2005 Jupiter Images Corporation.

THE BIG PICTURE:
Building Mosaics of the Entire Sky

SDSC Helps the National Virtual Observatory and Montage Stitch Together 10 Terabytes of 2MASS Images

by Paul Tooby

The old expression "can't see the forest for the trees" is just as apt for astronomers who observe the heavens as for viewers of earthly woodlands. As telescopes become ever more powerful, they give astronomers deeper views that reveal dramatic new details of narrow areas of the sky. But without a unified view of larger regions, scientists may miss large-scale patterns that can help them make sense of the individual objects they see.

What if astronomers had tools that let them see not only deeper into space but also broader areas of the sky—even the whole sky—in a unified image collection? That dream is now becoming a reality through advances in both telescopes and the supporting data cyberinfrastructure of software, data resources, and supercomputers being pioneered in a Strategic Applications Collaboration involving SDSC experts and astronomers from Caltech and NASA's Jet Propulsion Laboratory (JPL).

Using the Montage software, the researchers are stitching together into mosaics millions of individual images in the 2-Micron All Sky Survey, known as 2MASS, which contains an enormous 10 terabytes of data (10 terabytes is about 2,500 DVDs, around the size of the printed collection of the Library of Congress). A derived image product on this scale has never before been produced, and the resulting mosaics will let astronomers study structures as complete entities, opening the door to exploring previously unseen large-scale relationships in the Universe.

"You'll underestimate the true complexity of what's going on if you look only at narrow regions of the sky," said astronomer John Good of the Montage project at Caltech and JPL. "Having mosaics of the entire sky will help researchers construct 3-D maps of larger regions, giving us a much clearer understanding of how these structures evolved—we're finding that it's a very dynamic, explosive process."

For example, astronomers are finding that viewing the "big picture" can give them new insights into the complexities of large star formation regions that include a mixture of point sources, barely extended sources, and very complex extended structures. A large mosaic map of the region and its environs helps scientists understand the physical processes shaping it—the dynamic interplay of massive stars and their radiation fields with the surrounding medium of gas and dust. These powerful tools can help unravel the mysteries underlying the birth of new stars.

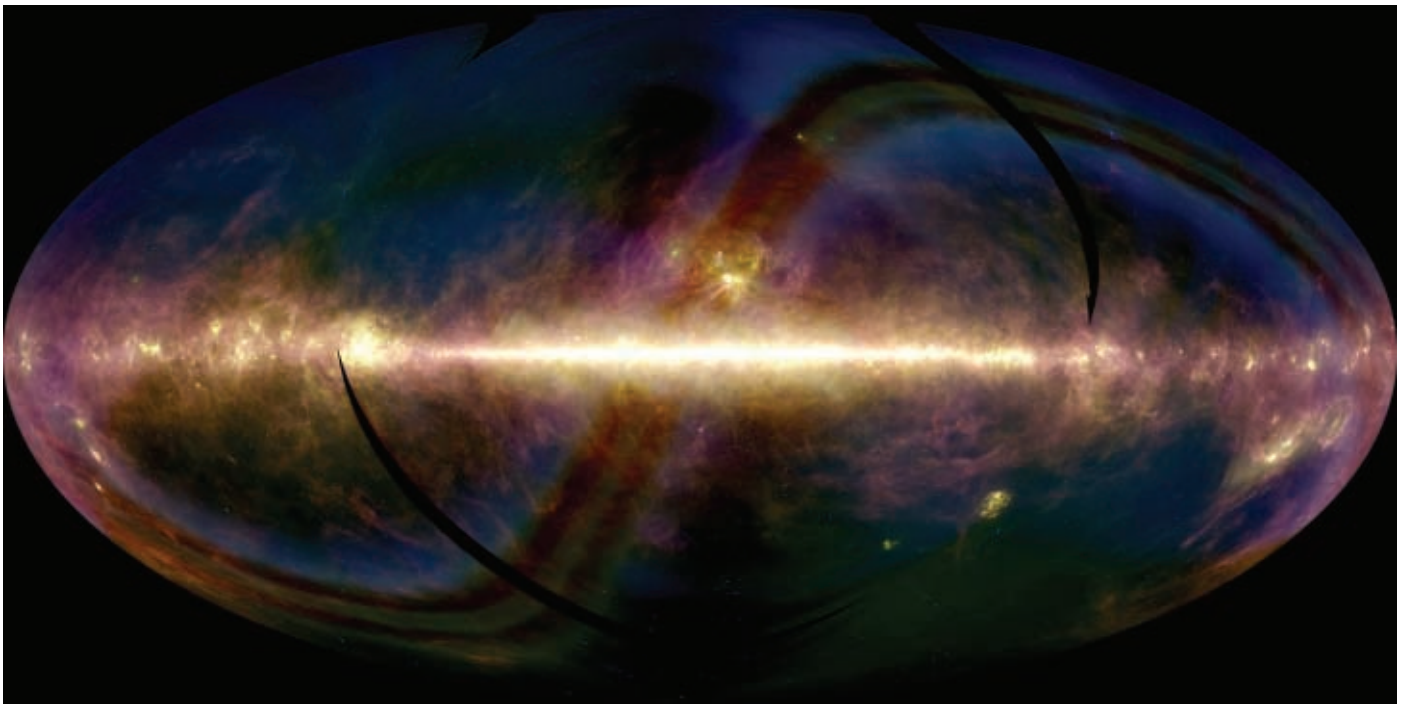
To produce the individual images in the 2MASS Survey, astronomers scanned the entire sky, using new instruments with an 80,000-fold improvement in sensitivity over earlier surveys. By looking at the sky in near-infrared light, the survey provides a "window" that is distinct from visible light, and can penetrate dust clouds of the Milky Way Galaxy to reveal an enormous number of objects that can't be seen in visible light. Funded by NASA and the National Science Foundation (NSF), the 2MASS survey is led by the University of Massachusetts, with

PROJECT LEADERS

Bruce Berriman, John Good, Thomas Prince, and Roy Williams, Caltech
Reagan Moore, SDSC/UCSD

PARTICIPANTS

Anastasia Clower Laity, Caltech
Atilla Bergou, Joseph Jacob, and Daniel Katz, JPL
Leesa Brieger and George Kremenek, SDSC/UCSD



Above: All-Sky Projection
An all-sky projection of the long-wavelength data from the Infrared Astronomical Satellite. The unusual form comes from the Aitoff projection, which maps the whole sphere into an equal-area projection. IPAC, Caltech.

Left: Milky Way Mosaic
This striking image of dust clouds in the plane of the Milky Way (running diagonally from upper left to lower right) is a three-color mosaic produced by SDSC and Montage from observations in the 2MASS all-sky survey. With 10,000 individual pixels on a side, the mosaic gives astronomers a view vastly expanded in both detail and width, opening the door to new insights into the large-scale structure of the Milky Way. Montage, IPAC, Caltech, SDSC/UCSD.

processing, construction, and distribution of the data products done by the Infrared Processing and Analysis Center (IPAC) at Caltech and JPL. Available through the IPAC Infrared Science Archive (IRSA), the 2MASS survey includes millions of high-resolution digital images.

In this collaboration, SDSC researchers are working with colleagues in the Montage project from IPAC and the Center for Advanced Computing Research (CACR) as well as the National Virtual Observatory (NVO) at Caltech and JPL.

THE WHOLE IS GREATER…

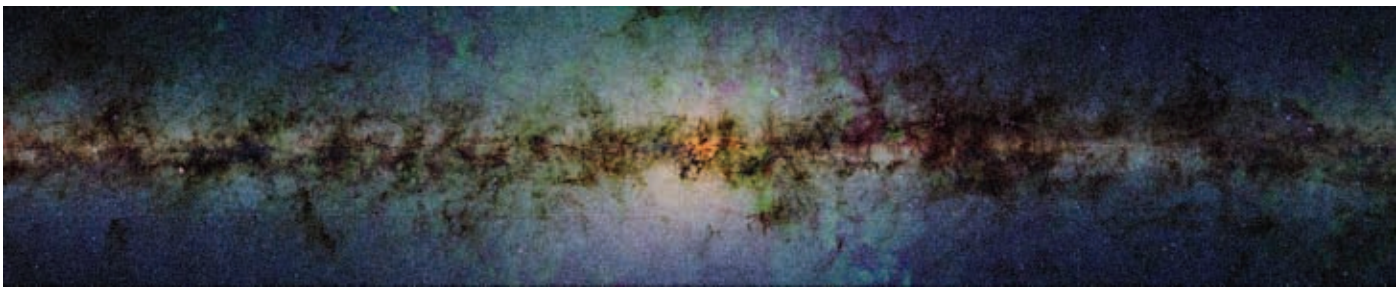
The Montage software being used to join individual 2MASS images can produce mosaics to user-specified parameters of projection, coordinates, size, rotation, and spatial sampling. The algorithms Montage uses are computationally intensive, and the 10 terabytes of data in the survey is huge. For this project the mosaic plates are 6 degrees on a side, and to cover the entire sky, SDSC staff scientist Leesa Brieger is generating 1,734 plates for each of the three near-infrared bands. Every 6 degree plate incorporates between 800 and 4,000 input files, with an average of about 1,500.

Brieger is working closely with the scientists in this Strategic Applications Collaboration to accelerate their research by developing ways to use large-scale SDSC resources most effectively. In addition to aiding this research project, the collaboration also aims to develop general solutions that will benefit the wider academic community. In processing the images, Brieger has developed an efficient workflow that first warps the input images onto the same projection, and then integrates the overlapping parts of adjacent images. Finally, a key task that Montage carries out is to apply a background model that rectifies to a common level the varying backgrounds of the millions of different images acquired at different times, under conditions with different background light, or sky emissions. The software does this by removing terrestrial and telescope features in the images in a way that can be tracked and described quantitatively.

The researchers anticipate that the project will take between 70,000 and 100,000 processor hours to complete on the TeraGrid at SDSC. The process is quite parallel, Brieger explains, and can be scaled up to run on more than 1,000 processors. Creating the mosaics is computationally demanding, requiring substantially more time than is needed to move the 10 terabytes of input data to SDSC and to store the expected 7 terabytes of output mosaics. The researchers are staging the input data on the local parallel file system and temporarily storing the output data locally until it is archived to join other long-term scientific data collections at SDSC as an important new

"Having mosaics of the entire sky will help researchers construct 3-D maps of larger regions, giving us a much clearer understanding of how these structures evolved—we're finding that it's a very dynamic, explosive process."

War and Peace Nebula
The region of star formation NGC 6357 is located about 7 degrees west of the Galactic center. This 2.5 degree square mosaic was constructed from near-infrared 2MASS and mid-infrared MSX images available through IRSA. Montage, Caltech.



Galactic Plane
Full-resolution mosaic of 2-Micron All Sky Survey (2MASS) galactic plane data constructed with Montage. IPAC, Caltech.

resource for astronomers.

Managing the mosaicking process with such massive amounts of data is presenting interesting new challenges. For example, with today's state-of-the art hardware and file systems, the data retrieval error rates are very low, on the order of one error in every trillion bytes. With such low error rates, an astronomer computing a single mosaic with Montage using a few thousand images typically won't encounter any problem.

However, as SDSC tackles the larger challenge of scaling up to compute thousands of mosaics for the entire 10 terabyte survey, even this tiny error rate can become significant. "When dealing with a multi-terabyte data collection with millions of files you can't ignore even small error rates," said Brieger. "The real contribution we make at SDSC is to help identify the problems that show up as scientists scale up their data and computing, and develop robust, fault-tolerant tools that let them get science done at the very large scale required at the frontiers of today's science."

A key resource for the project is SDSC's massive data storage resources—more than one petabyte (one million gigabytes) of online disk backed by six petabytes of archival tape, along with the SDSC Storage Resource Broker (SRB) software for organizing, moving, and archiving massive data collections. "SDSC's end-to-end data and computational capabilities are enabling the astronomy community to reach the milestone of being able to process the entire 10 terabyte 2MASS sky survey," said Reagan Moore, Distinguished Scientist and director of the Data Intensive Computing Environments (DICE) group at SDSC. "The 2MASS survey is archived using the SRB in SDSC's HPSS archival storage, and with the help of SDSC expertise and resources, the Montage software is able to produce mosaics of the entire survey."

AN ATLAS OF THE ENTIRE SKY

An important step, the researchers explain, is to validate the mosaics against the original 2MASS survey images from which they were derived. The completed science-grade mosaics are then stored along with information on all of the modifications of the raw images, tracked in a quantitative way so that astronomers can know exactly how each plate was produced. And the mosaics are being made in accordance with Hyperatlas standards and will form part of the growing Hyperatlas coordinated by Roy Williams of Caltech, which will be available for public access through the NVO.

By producing a collection of uniform mosaics for the entire 2MASS all-sky survey, the collaboration represents an important step forward for astronomers. And when the large-scale mosaic project is extended so that a number of surveys in different wavelengths are re-projected to the same pixel grid, or map projection, then images from surveys in different wavelengths can be directly overlaid and jointly compared and analyzed in ways never before possible.

This is exciting, explains Williams, because it opens the door for large-scale data mining of the "federated pixels," that is, the background-adjusted mosaicked images that are linked through a common catalog and available in a common projection. Astronomers will be able to ask complex, "big picture" questions that both explore the different wavelengths of various surveys and extend across large spatial structures in the Universe—a new mode of inquiry.

In addition to the 2MASS survey, Caltech researchers are re-projecting the 3 terabyte Digital Palomar Observatory Sky Survey collection, also stored in a SRB collection at SDSC, and preparing to do the same with the Palomar-Quest survey, already 13 terabytes in size and growing by one terabyte a month. Eventually the Sloan Digital Sky Survey (SDSS) may form part of the collection as well.

Collaborators in the project include Thomas Prince, Montage Principal Investigator, Bruce Berriman, Montage project manager, Anastasia Clower Laity, John Good, and Roy Williams, NVO co-director, of Caltech; Joseph Jacob, Daniel Katz, and Atilla Bergou of JPL; and Leesa Brieger, George Kremenek, and Reagan Moore of SDSC. Montage is funded by the NASA Earth Science Technology Office, Computational Technologies Project.

– Paul Tooby is a senior science writer at SDSC and editor of EnVision Magazine.

"The real contribution we make at SDSC is to help identify the problems that show up as scientists scale up their data and computing, and develop robust, fault-tolerant tools that let them get science done at the very large scale required at the frontiers of today's science."

RELATED LINKS

National Virtual Observatory (NVO)
www.us-vo.org

Montage
montage.ipac.caltech.edu

SDSC Strategic Applications Collaborations (SAC)
www.sdsc.edu/user_services/sac/index.html

SDSC News

SDSC EXPANDS DATASTAR SUPERCOMPUTER TO 15.6 TERAFLOPS

To better support the extreme data-intensive needs of U.S. science and engineering researchers and educators, SDSC has expanded the capacity and capability of its DataStar supercomputer to more than 2,048 processors. Through the addition of 96 8-way IBM Power 4+ p655 compute nodes, users will now have access to one of the most powerful computers in the nation available to the open academic community.

The expanded DataStar will provide SDSC users 50 percent more capacity, helping meet the heavy demand for the center's compute time. In addition, the aggregate memory as well as the size of DataStar's parallel file system will almost double. This will provide 7.3 terabytes of aggregate memory and 115 terabytes of parallel file system disk storage, giving researchers the ability to compute and output more data in research areas such as astronomy, geosciences, fluid dynamics, and others. SDSC is the TeraGrid site with particular responsibility for data-intensive computing, and the expansion will greatly improve overall performance, providing a more powerful tool for users.

"Data-intensive computing has rapidly become a principal mode of scientific exploration. This expansion of SDSC's DataStar system demonstrates SDSC's and the NSF's recognition that the best tools must be made available to the science and engineering communities," said José Muñoz, deputy director of the Office of Cyberinfrastructure at the National Science Foundation (NSF). "SDSC has been a leader in data-intensive computing, and this upgrade maintains that leadership. The capabilities being made available through this expansion, coupled with SDSC's excellent scientific and support staff, are paramount for continued U.S. leadership in science, engineering, and education."

DIRECTOR BERMAN CO-CHAIRS NSF SBE-CISE WORKSHOP

The National Science Foundation funded the Social, Behavioral, and Economic Sciences/Computer and Information Science and Engineering (SBE/CISE) Workshop on "Cyberinfrastructure for the Social and Behavioral Sciences" in recognition of NSF's role in enabling, promoting, and supporting science and engineering research and education. The workshop was designed to help social, behavioral, and economic researchers identify their needs for infrastructure, their potential for helping CISE develop this infrastructure for engineering and all the sciences, and their capacity for assessing the societal impacts of cyberinfrastructure. SDSC Director Fran Berman co-chaired the workshop along with Henry Brady of the University of California, Berkeley. More than 80 leading CISE and SBE scientists were brought together at Airlie House in Virginia on March 15-16, 2005 to discuss six areas: cyberinfrastructure tools for the social and behavioral sciences; cyberinfrastructure-mediated interaction; organization of cyberinfrastructure and cyberinfrastructure-enabled organizations; malevolence and cyberinfrastructure; the economics of cyberinfrastructure; and finally, the impact of cyberinfrastructure on jobs and income. For more information and the final workshop report see http://vis.sdsc.edu/sbe/.

LIBRARY OF CONGRESS AND NSF ANNOUNCE DIGARCH DIGITAL PRESERVATION AWARDS

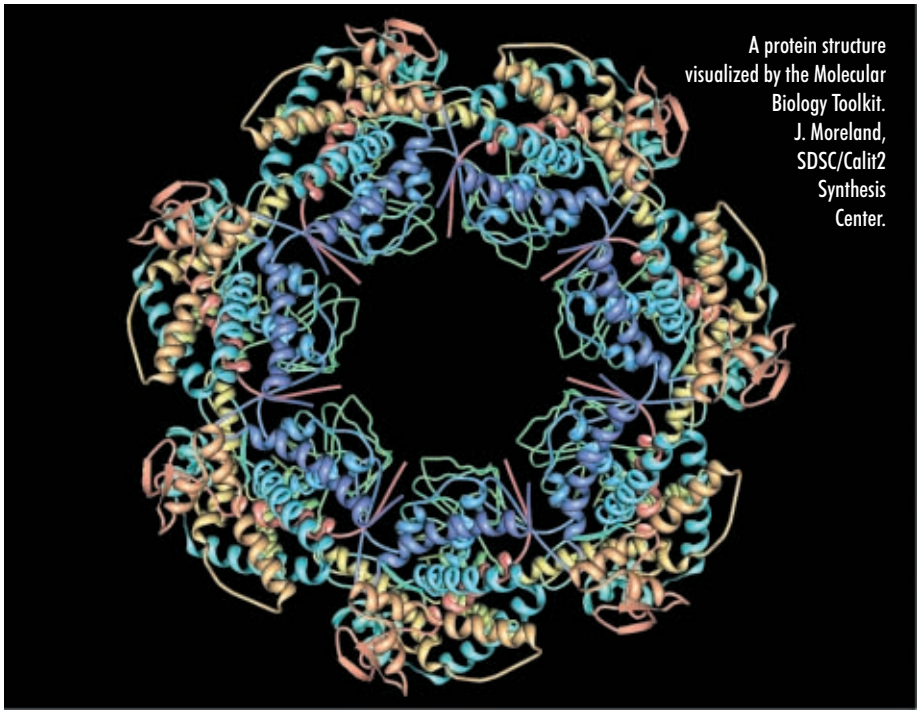
The Library of Congress National Digital Information Infrastructure and Preservation Program (NDIIPP) and the National Science Foundation awarded 11 university teams a total of $3 million to undertake pioneering research to support the long-term management of digital information. These awards are the outcome of a partnership between the two agencies to develop the first digital-preservation research grants program.

The Library is implementing a national digital preservation strategy, which involves building a collaborative network of partners to collect at-risk content and to develop new preservation tools. Research supported under these awards will help produce the technological breakthroughs needed to keep very large bodies of digital content securely preserved and accessible over many years.

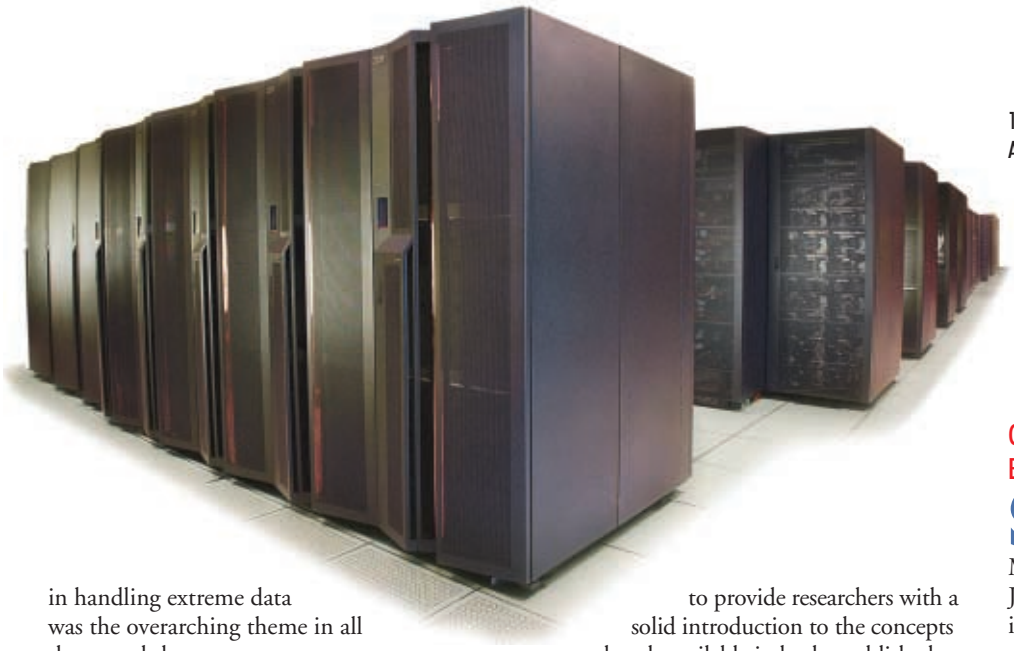
SDSC, a world leader in preservation technologies, will participate in the Digital Preservation Lifecycle Management project, demonstrating this process for video content. Researchers will develop and document a practical preservation process for mixed collections of both legacy and "born digital" video material. SDSC will also participate with the Scripps Institution of Oceanography and the Woods Hole Oceanographic Institution in the Multi-Institution Testbed for Scalable Digital Archiving to develop a multi-terabyte digital repository to preserve data from more than 1,600 oceanographic research projects.

SDSC HOSTS THREE SUMMER WORKSHOPS

SDSC's in-demand educational activities included hosting three summer workshops—its first Blue Gene Users Workshop, the second annual GEON Cyberinfrastructure Summer Institute, and the 11th Annual Computing Institute for Scientists and Researchers. SDSC's expertise



A protein structure visualized by the Molecular Biology Toolkit. J. Moreland, SDSC/Calit2 Synthesis Center.



15.6 teraflops SDSC DataStar supercomputer. A. Decker.

in handling extreme data was the overarching theme in all three workshops:

- The Blue Gene Users Workshop, held July 7–8, 2005 at SDSC, emphasized hands-on access for 18 researchers from a variety of scientific fields who require the large number of processors on SDSC's newest supercomputer. The Blue Gene system, with 2,048 compute processors in a single rack, is nicknamed "Intimidata" because it is specially configured with 128 I/O nodes to support data-intensive computing. The hands-on sessions were complemented by lectures from experts at SDSC and the Lawrence Livermore National Laboratory, home of the world's largest Blue Gene system.
- Thirty-eight researchers attended the second annual GEON-hosted Cyberinfrastructure Summer Institute for Geoscientists (CSIG). The attendees included graduate students, postdocs, and researchers in geoscience and information technology from a number of agencies and more than 30 institutions around the U.S. and as far away as Japan, Korea, and the United Kingdom. The popular one-week educational program, held from July 18-22, gives geoscientists an "IT headstart" in using powerful new information technologies, or cyberinfrastructure, to enable a new generation of geoscience discoveries.
- The 11th Annual Computing Institute at SDSC hosted 30 researchers for a week-long lecture series and hands-on laboratory focusing on managing large data sets with "extreme I/O." Attendees included graduate students, scientists, and researchers representing 17 institutions in the U.S. and abroad. The program, held July 25-29, was designed to provide researchers with a solid introduction to the concepts and tools available in both established and new technologies for the creation, manipulation, dissemination, and analysis of large data sets.

SDSC LAUNCHES 'DATA CENTRAL'

Interdisciplinary collaborations and community-shared data collections and databases are becoming increasingly important to the progress of science. For this reason, SDSC launched Data Central, which allows users to make their data collections and databases publicly available to a wide community of users. The site also offers hosting and long-term archiving that provide researchers with easy-to-use and complete data management services, freeing them to focus on their research. With storage facilities of more than one petabyte (1,000 terabytes) of online disk and six petabytes of archival tape storage, SDSC currently hosts more than 50 publicly available data collections from bee behavior videos and multi-terabyte all-sky astronomy image collections to data from the Library of Congress.

Eligible researchers can request a data allocation (whether or not they have a compute allocation) that permits expanded access to SDSC's facilities for data collection hosting, database hosting, and long-term archiving. To request a data allocation, fill out the simple, fast online form, which can be submitted automatically online. To learn more, see the Data Central site at http://datacentral.sdsc.edu/.

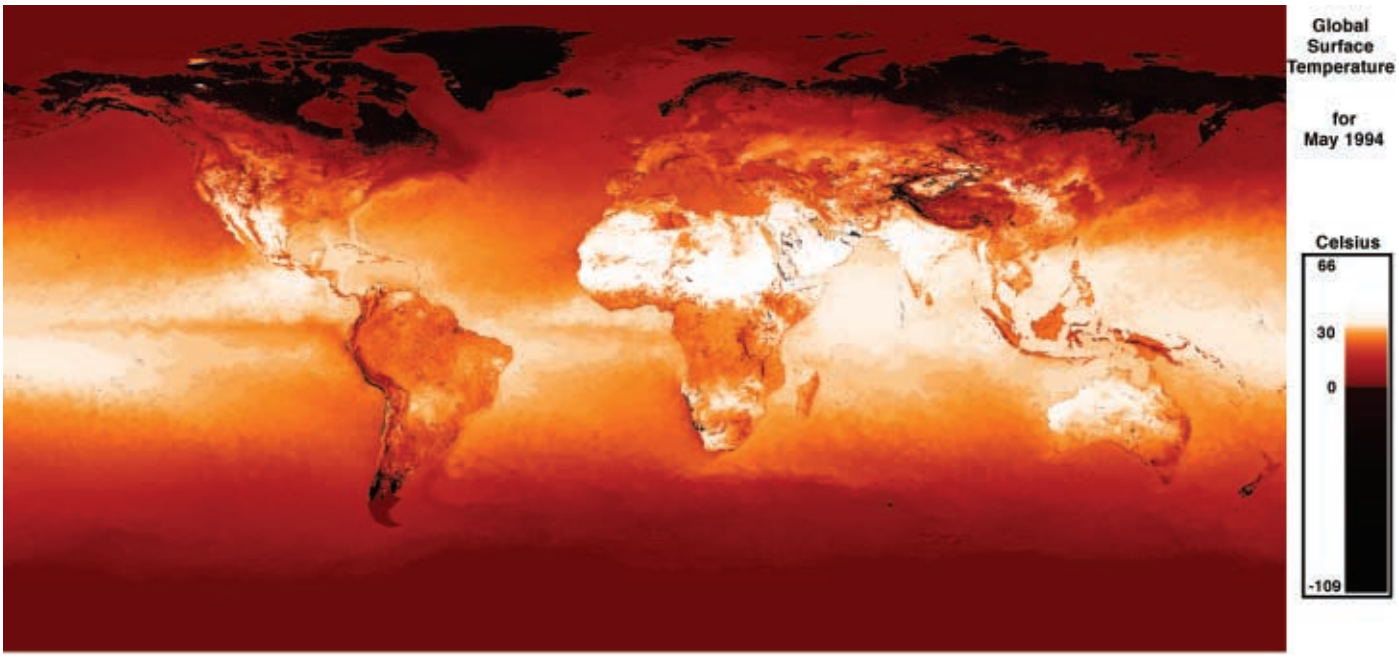
CENTER INTRODUCES MOLECULAR BIOLOGY TOOLKIT

SDSC has introduced the National Institutes of Health (NIH)-funded Molecular Biology Toolkit (MBT), a set of Java-based software libraries for manipulating, analyzing, and visualizing information about proteins, DNA, and RNA. This first major release of the MBT runs under the Linux, Windows, Mac OS X, and IRIX operating systems—an important advantage since very few off-the-shelf packages enable applications to run seamlessly on several different computer platforms. The MBT includes source code, example applications, a Programmer's Guide, an Application Program Interface (API) document, a Build Guide, and a Binary Installation Guide.

The toolkit provides Java classes for efficiently loading, managing, and manipulating protein structure and sequence data. The MBT also provides a rich set of graphical 3-D and 2-D visualization modules which can be plugged together to produce applications that have sophisticated graphical user interfaces. And the core data I/O and manipulation tools can also be used to write completely non-graphical applications—to implement pure analysis codes, for example, or to produce a non-graphical back end for Web-based applications. More information is at http://mbt.sdsc.edu/.

Thirty researchers attend 11th Annual Computing Institute held this summer at SDSC. B. Tolo.

SDSC SAN DIEGO SUPERCOMPUTER CENTER



LONG-TERM RECORDS OF GLOBAL SURFACE TEMPERATURE

This visualization of global surface temperature is part of a National Science Digital Library collection archived at SDSC. Preservation technologies developed at SDSC are capable of maintaining educational and scientific data for decades or more. This is essential to scientists' efforts to understand and predict important phenomena such as global temperature and climate change, and to enable the NSDL to educate the next generation of researchers. More information on the NSDL is at http://nsdl.org/. NSDL collection support C. Cowart, SDSC/UCSD; image created by G. Shirah and J. Kermond, NASA GSFC Scientific Visualization Studio.

SUBSCRIPTIONS

For a free subscription to enVision, send the information requested below to:

Gretchen Rauen, SDSC
University of California, San Diego
9500 Gilman Drive, MC 0505
La Jolla, CA 92093-0505

eradmin@sdsc.edu, 858-534-5111

www.sdsc.edu

NAME

TITLE

INSTITUTION

ADDRESS

CITY, STATE, AND ZIP

COUNTRY

E-MAIL

SDSC

University of California, San Diego
San Diego Supercomputer Center
9500 Gilman Drive, MC 0505
La Jolla, CA 92093-0505

ADDRESS SERVICE REQUESTED

NON-PROFIT ORG
U.S. POSTAGE
PAID
WESTERN GRAPHICS