NPACI & SDSC

Quarterly Science Magazine

EnVision

VOL. 19 NO. 1

JANUARY – MARCH 2003

The Encyclopedia of Life to Open New Chapter of Biological Discovery

Mining the Heavens: the National Virtual Observatory

Preserving Priceless Memories for the Library of Congress

Sapphire Internet Worm Shatters Speed Records

NPACI

Building the Computational Infrastructure for Tomorrow's Scientific Discovery

The National Partnership for Advanced Computational Infrastructure (NPACI) joins 41 partner institutions in 17 states, Australia, Italy, Spain, and Sweden, in creating the foundation for a ubiquitous, continuous, and pervasive computational environment to support research by the world's scientists. NPACI is led by UC San Diego, funded primarily through the NSF's Partnerships for Advanced Computational Infrastructure program, and has its focus of activities at the San Diego Supercomputer Center (SDSC).

NPACI INFORMATION

Fran Berman, Director

Bart McDermott, Director of Development and Communications

SDSC

SDSC is a campus research unit of UC San Diego and the focus of activities for NPACI. Since 1985, SDSC has served the U.S. scientific community as a national laboratory for computational science and engineering.

GENERAL INFORMATION

Phone: 858-534-5000

info@npaci.edu

www.npaci.edu

www.sdsc.edu

Subscribe to enVision and *Online* at: www.npaci.edu/Press/subscriptions.html

ONLINE

www.npaci.edu/Online

Online is a biweekly, Web newsletter of the latest research and developments from NPACI and SDSC.

enVision

www.npaci.edu/enVision

enVision, ISSN 1521-5334, is published quarterly by NPACI and SDSC Communications. For a free subscription or to make address changes, visit the website or contact the editor.

EDITOR: Cassie Ferguson
cferguson@sdsc.edu
858-534-5000

DESIGNER: Gail Bamber
bamberg@sdsc.edu
858-534-5150

Contents

ERRATUM

In the third quarter issue of EnVision, in a story titled, "GridPort Provides Simple Entrance to Scientific Computing," the funding source for the National Biomedical Computation Resource (NBCR) program was erroneously mentioned as being the National Science Foundation. In fact, the NBCR is supported by the National Institutes of Health through a National Center for Research Resources program grant to researchers at the University of California, San Diego, including the San Diego Supercomputer Center and The Scripps Research Institute. The mission of the National Biomedical Computation Resource is to conduct, catalyze, and enable biomedical research by harnessing advanced computational technology.

MONTAGE PROJECT AT CACR, JPL, AND IPAC, CALTECH

FRONT COVER: Infrared Mosaic of a Section of Our Galaxy

A three-color mosaic derived from images in the Second Incremental Data Release of the 2 Micron All Sky Survey (2MASS). The picture shows dust clouds and nebulosity in the plane of our Galaxy, and it combines images at three near-infrared bands. The mosaic contains 12,000 individual pixels on a side. There are 347 individual images in each of three bands: J (1.25 μm), H (1.65 μm), and K (2.2 μm). Montage is funded at the Caltech and JPL through a contract with NASA's Earth Sciences Technology Office Computational Technologies Project. Technical details are at the website: montage.ipac.caltech.edu. For information on 2MASS see www.ipac.caltech.edu/2mass/*

*This product is not endorsed by the 2MASS project, and is intended as a proof of concept of the Montage algorithm rather than a science product.

Building a Successful Cyberinfrastructure

HARRY AMMONS, SDSC

BY FRAN BERMAN
NPACI and SDSC Director

Last December, the National Science Board (NSB) released a draft of a comprehensive report on infrastructure for U.S. science and engineering, *Science and Engineering for the 21st Century: The Role of the National Science Foundation*. The NSB report placed a high priority on substantive investment in a comprehensive, competitive, and broad-based U.S. infrastructure, stating, "The Nation's IT capability has acted like adrenaline to all of S&E. The next step is to build the most advanced research computing infrastructure while simultaneously broadening its accessibility."

At the beginning of February, the National Science Foundation (NSF) released the long-awaited Blue Ribbon Advisory Panel report on cyberinfrastructure. The report, entitled *Revolutionizing Science and Engineering through Cyberinfrastructure,* is broad in scope and will have enormous impact on the next generation of science and technology programs and activities at NSF, as well as in the larger community. The Blue Ribbon Panel report presents a compelling vision of the next generation of cyberinfrastructure. Its main recommendations include advocating the development of a multisite, comprehensive cyberinfrastructure program, funded at over $1 billion per year, which will support a full spectrum of coordinated information technology activities, including hardware deployment, networking, grid technologies, software development, data centers, education, documentation and training, disciplinary outreach, and science and technology research.

As envisioned in both reports, cyberinfrastructure will require an unprecedented coordination of hardware, software, and human infrastructure to meet the needs of science and society and to leverage the increasingly ubiquitous and sophisticated technological landscape of the modern world.

HOW WILL WE BUILD A SUCCESSFUL CYBERINFRASTRUCTURE?

Cyberinfrastructure provides a once-in-a-generation opportunity to leverage the extraordinary momentum of the science and technology community to build a global information infrastructure. The NSF program has more than a decade of hands-on experience developing and using computational and data management infrastructure to enable new generations of scientific advances. This is a defining moment. *The program NSF will now initiate has the potential to change the direction and evolution of science and technology.* The program's design, goals, mission, and structure will have a fundamental impact on our scientific culture. To achieve cyberinfrastructure's potential, I believe that the new program must be founded on the following principles:

- **Cyberinfrastructure must incentivize for real cooperation among its participants.** The value of cyberinfrastructure is its ability to provide end-to-end solutions. Such solutions will require close cooperation among cyberinfrastructure participants and the ability to plan and execute a well-thought-out, well-designed, useful infrastructure that can be deployed in the short term and evolved over the long term.
- **Cyberinfrastructure must incentivize for serious software infrastructure development.** At its heart, cyberinfrastructure is predicated upon a complex, integrated, interoperable, distributed system that must be robust, stable, persistent, evolutionary, and usable. *Cyberinfrastructure must support large-scale, daily use of its resources, and support science at all levels, not just demos.* All cyberinfrastructure software will benefit from software and usability engineering, and users will derive the maximal impact from substantive cyberinfrastructure operations support, documentation, outreach, and training activities. Many of the challenges in developing successful, working cyberinfrastructure will require fundamental and targeted research for which the most promising ideas must be coupled with prototyping, testing, and implementation efforts.
- **Cyberinfrastructure should provide a full spectrum of resources to be of maximal benefit.** The promise of cyberinfrastructure is that it will become our national computational and data management fabric, able to support a broad spectrum of activities from simple access to the largest-scale computations, with the potential to enable breakthrough science advances. Large-scale resources must be available and coupled with smaller scale resources to provide an evolutionary path from development to large-scale execution. Data and computation resources must be linked to support analysis and knowledge synthesis of the immense and increasing deluge of data. *Cyberinfrastructure must provide a balanced portfolio of resources, from the low end to the high end, as well as robust, usable software that allows users to integrate these resources together.*
- **Cyberinfrastructure must include disciplinary scientists, social scientists, computer scientists, and technologists.** For the last two decades, interdisciplinary teams of domain scientists and computer scientists have addressed some of the most fundamental problems in scientific research. Domain scientists must be an integral part of cyberinfrastructure as the requirements of science and engineering applications will both drive and inform cyberinfrastructure development. Social scientists and economists will be needed to help address the complex policy and "market" aspects of cyberinfrastructure. Finally, computer scientists and technologists must be on board to address the enormous challenges of designing, developing, and deploying the complex systems that will enable cyberinfrastructure. All of these

activities will require research, experimentation, implementation, and coordination.

- **The scope, goals, structure, and budget of the cyberinfrastructure program must be aligned.** It is critical that the participants of cyberinfrastructure are expected and enabled to deliver on cyberinfrastructure's promise. *This means that careful attention must be paid to aligning the goals of the program, the scope of its activities, and the costs of delivering on its promises.* Achieving a petaflop in computer power, enabling results that will appear in distinguished journals such as *Science* or *Nature*, and increasing the number of graduate and undergraduate degrees in the sciences and technology are all distinct and important goals which may require different organizational, technical, and fiscal strategies. The cyberinfrastructure program must clearly identify, prioritize, and be structured to support its goals.
- **The cyberinfrastructure program must account for the human infrastructure required for development, deployment, and operational success.** Cyberinfrastructure must recognize the importance of human infrastructure. Individuals with multidisciplinary, science, and technology expertise are key to the process of designing, developing, and evolving cyberinfrastructure, and will prove to be cyberinfrastructure's most valuable component.

To evolve a broad-based, high-impact, successful national information infrastructure that can achieve its immense potential, the cyberinfrastructure program will need goals, metrics, timeframes, and an organizational and management structure that will promote and ensure its success. The science and technology community and the PACI partnerships stand ready to assist the NSF in designing such a program, and are committed to addressing the challenges of developing a successful and working cyberinfrastructure that achieves the vision of the national information infrastructure so compellingly described in the National Science Board and Blue Ribbon Panel reports. ▼

NSF Report Envisions a Cyberinfrastructure That Will Empower Science and Engineering

The critical needs of science and rapid progress in information technology are converging to provide a unique opportunity to create and apply a sustained cyberinfrastructure that will "radically empower" scientific and engineering research and allied education, according to the National Science Foundation (NSF)'s Advisory Committee for Cyberinfrastructure. The committee details its recommendations in a recently released report, entitled *Revolutionizing Science and Engineering through Cyberinfrastructure.*

Like the physical infrastructure of roads, bridges, power grids, telephone lines, and water systems that support modern society, "cyberinfrastructure" refers to the distributed computer, information and communication technologies combined with the personnel and integrating components that provide a long-term platform to empower the modern scientific research endeavor.

Cyberinfrastructure is "essential, not optional, to the aspirations of research communities." For scientists and engineers, the report states, cyberinfrastructure has the potential to "revolutionize what they can do, how they do it, and who participates." The seeds of this revolution are seen in community-driven efforts, supported by NSF and other agencies, such as the Network for Earthquake Engineering Simulations (NEES), the Grid Physics Network (GriPhyN) and the National Virtual Observatory (NVO).

"We've clearly documented extensive grass-roots activity in the scientific and engineering research community to create and use cyberinfrastructure to empower the next wave of discovery," said Dan Atkins, chair of the advisory committee and professor in the University of Michigan School of Information and the Department of Electrical Engineering and Computer Science. "We're at a new threshold where technology allows people, information, computational tools, and research instruments to be connected on a global scale."

While identifying the opportunities, the committee warned that the cyberinfrastructure that is needed cannot be created today with off-the-shelf technology. As a result, they called for increased fundamental research in computer science and engineering.

In addition to NSF's support for projects such as NEES, GriPhyN and NVO, the report calls out NSF's leadership in the Partnerships for Advanced Computational Infrastructure (PACI) program, the TeraGrid effort, the NSF Middleware Initiative (NMI), the Digital Libraries Initiative and the Information Technology Research program as providing a foundation for the future cyberinfrastructure.

Its unique breadth of scientific scope and prior investments position NSF to lead an interagency program to develop an advanced cyberinfrastructure for the nation, according to the report. To reach critical mass, an advanced cyberinfrastructure activity would require interagency partnerships as well as collaboration between the physical and life sciences, computer science, and the social sciences.

The opportunity is evidenced by both progress from developments in information technology and the mushrooming of cyberinfrastructure projects for specific fields, initiated by scientists in those fields. The NSF has a "once-in-a-generation opportunity," according to the committee, to lead the scientific and engineering community in the coordinated development and expansive use of cyberinfrastructure.

—David Hart, NSF ▼

Encyclopedia of Life to Open New Chapter of Biological Discovery

When researchers published the first draft of the human genome, the media hailed the accomplishment as one of the most significant achievements in modern science. In the surrounding years, scientists have sequenced the genomes for more than 800 additional organisms, from flies to mice, rice to corn. Yet, the growing catalog of genomes has led the scientific community has begun to ask: What's next?

Within an organism, the DNA that forms a genome is translated into proteins, which are then modified to perform specific functions such as carrying messages or buttressing cell walls. Theoretically, knowing an organism's genetic sequence will reveal the structure and function of every protein it produces. However, proteins are notoriously complex molecules and deciphering their structures and functions remains one of the foremost challenges of biology. A system that could automatically marry the enormous amount of genomic data with the latest knowledge of proteins would provide a sequel to the genome projects.

Scientists in an open collaboration led by the San Diego Supercomputer Center (SDSC) are building such a system, the Encyclopedia of Life (EOL). The EOL is an ambitious project to catalog the complete set of proteins, or proteomes, of every living species in a flexible, powerful reference system available via the Web. The project not only draws on the skills of biologists but also experts in high performance computing, data and knowledge systems, and grid computing, as well as of the most powerful computational resources ever to be applied to a biological problem, NPACI's Blue Horizon and, ultimately, the TeraGrid.

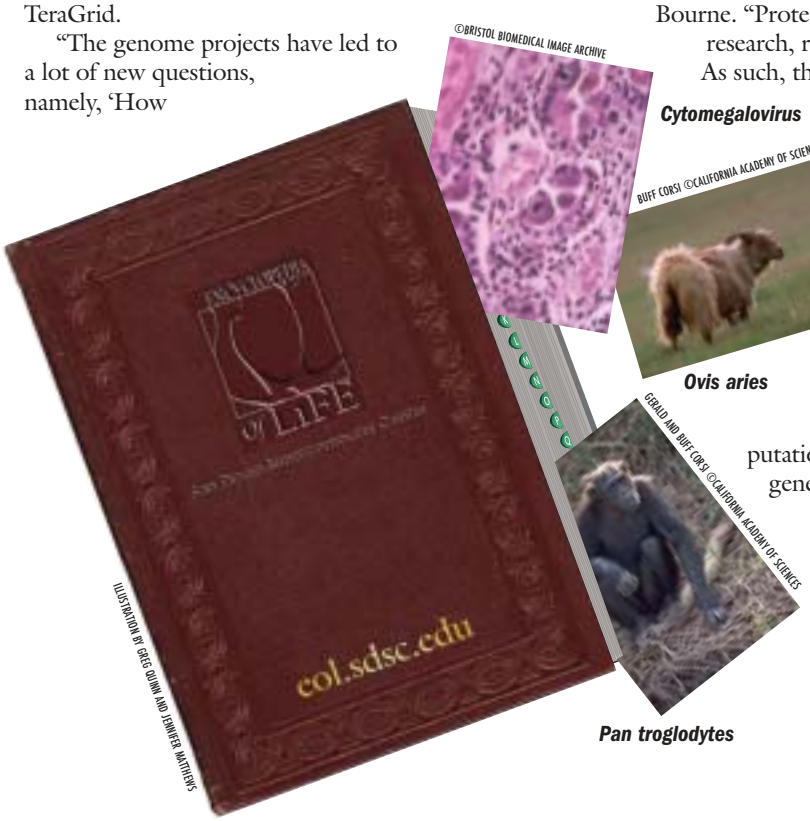
"The genome projects have led to a lot of new questions, namely, 'How can we use this sequence information?'" said Philip Bourne, SDSC's director of integrative biosciences and UCSD professor of pharmacology. "The Encyclopedia of Life is part of the answer to those questions. The EOL will point out attractive drug targets and identify domain boundaries. It will compare proteins from multiple species, or even all the sequenced species available in the EOL repository."

The EOL will be used by scientists in areas such as biomolecular-based therapeutics, industrial enzymes, bacterial-based bioremediation, and counter-bioterrorism. Using the EOL, scientists will be able to uncover the prevalence of a given protein across all kingdoms of life, molecular interactions with that protein, and whether the function of the protein varies across species. The EOL will also serve as an educational tool that caters both to undergraduates who might be investigating a particular organism to elementary school students who might be learning about proteins for the first time.

THE *ENCYCLOPÆDIA BRITANNICA* OF THE LIFE SCIENCES

"The Encyclopedia of Life strives to be the *Encyclopædia Britannica* of the life sciences," said Bourne. "Proteins are central to all biological research, regardless of the scale of the work. As such, this is the era of the protein and the EOL will endeavor to be the basis for studies in areas such as protein-protein interactions, biochemical pathways, and, ultimately, systems biology."

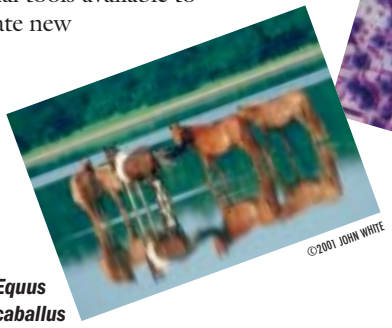
Slated for public release later this year, the EOL will be a massive database that combines all that is known about proteomes with the most powerful computational tools available to generate new



Cytomegalovirus

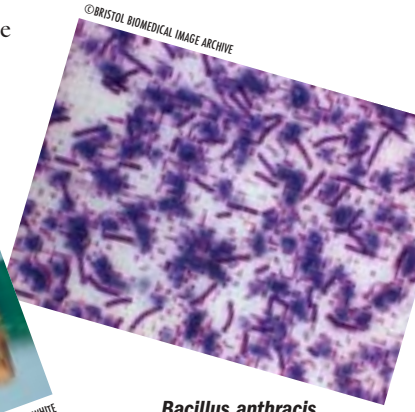
Ovis aries

Pan troglodytes



Equus caballus

Drosophila melanogaster



Bacillus anthracis

PROJECT LEADERS

Philip Bourne, Mark Miller
SDSC

PARTICIPANTS

Kim Baldridge, Chaitanya Baru, Fran Berman, Philip Bourne, Robert Byrnes, Henri Casanova, Neil Cotofana, Tony Fountain, Jerry Greenberg, Wilfred Li, Coleman Mosley, Dmitry Pekurovsky, Greg Quinn, Vicente Reyes, Peter Shin, Ilya Shindyalov, Stella Veretnik
SDSC

eol.sdsc.edu

understandings of diseases and the natural world. Such data is a fundamental part of drug discovery and agricultural engineering and provides insights that are not detectable from protein sequence analysis alone.

"Our goal is to be the storehouse and portal for proteomics worldwide. Proteomics includes not only the identification and quantification of proteins, but also the determination of their localization, modifications, interactions, activities, and, eventually, their function," said Mark Miller, EOL project manager. The EOL will incorporate data from a number of other protein-related projects at institutions worldwide.

Once the EOL contains a critical volume of species, scientists will be able to solve problems that currently require an extraordinary amount of time, computing resources, and money to investigate. The EOL will perform new, powerful comparisons of 3-D protein structures across multiple species. It will perform predictions that improve understanding of the mechanism of protein function, such as whether a protein that adopts a certain kind of fold will have the same active site as other similarly shaped proteins.

The EOL will offer an easy-to-use Web repository of high-quality data and tools for virtually any matter related to proteomic information. All of the entries in the EOL will be integrated with other biological resources and databases, answering questions such as:

- Is protein X found in anthrax?
- What other species contain protein X?
- Has caspase-1, a protein involved in mammalian cell death and aging, been identified in any plants and do the protein structures in the different species look similar?
- If altering a gene in *Arabidopsis* leads to a 10 percent increase in growth, will the same response happen in rice?
- What protein folds appeared early in evolution?
- Do disease genes already identified in humans occur in EOL with different putative structures and functions that those previously assigned?
- Protein folds are reused to perform multiple functions and the same function may be performed by multiple folds. How does this many-to-many relationship change across species?
- A potential drug target is identified in humans. What other model organisms contain the same protein with the same structure and function and are amenable to *in vivo* experiments such as gene knockouts?

Xenopus Laevis

Canis familiaris

Rattus norvegicus

According to Bourne, the number of genomes available to the public is increasing so rapidly that within 10 years, people will have access to their personal DNA sequence as a tool for medical diagnosis. However, that genomic data only provides a starting point for further action. The EOL will offer the resources to build on genomic discovery.

However, accessing the data needs to be easy, and the EOL team has emphasized the importance of usability in its interface and in a comprehensive suite of tools that will be used to browse the EOL.

During a demonstration of the EOL given at SC2002, the annual conference on high performance computing and networking held last November, Greg Quinn, software lead on the EOL project and senior programmer analyst at SDSC, explained the computational aspects of the project to an audience consisting of researchers in the high performance computing community. He pointed out the methods that would be used to generate the data as well as the tools that would be available to biologists, who often do not have the technical know-how required to navigate a complex database.

In an effort to make the EOL as easy-to-use as possible, Quinn has been developing an intuitive interface, literally based on the idea of accessing the EOL like a book. This book metaphor makes extensive use of scalable vector graphics data rendering components that the end-user can interactively query in real-time for annotation data. The EOL group is also designing a client-side application, the EOL Notebook for storing and automatically updating data via EOL Web services, and a software package that will be distributed for deploying the EOL data and delivery mechanism at other research centers.

"Creating an intuitive interface and powerful search methods for the massive amounts of EOL data is essential," said Quinn. "We are engineering the system from the ground up to support multiple methods of data access by researchers, through both an intuitive Web interface and SOAP-based data retrieval methods. XML data download is also available from search and BLAST results for advanced users of the system. We feel the EOL Notebook is an essential component of this model, enabling the end user to store and collate the potentially large amounts of data retrieved during a user session, with the EOL SOAP Web services automatically keep that locally stored data current."

Conus sp.

Arabidopsis thaliana

RECORD-BREAKING BIOLOGICAL COMPUTATION

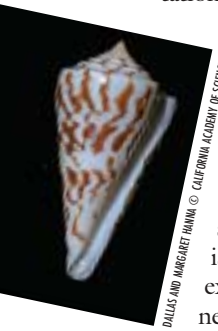
The first step in filling the volumes of the EOL involves transforming genomic information into proteomic information by running it through a pipeline—a fully automated process for running a series of programs developed at SDSC and elsewhere. Late in 2001, Bourne, working with Ilya Shindyalov, group leader in protein science research at SDSC, Wilfred Li, senior programmer analyst at SDSC , and the company Ceres Inc., designed such a pipeline to transform, or annotate, the genome of a plant called *Arabidopsis thaliana*.

The pipeline, called iGAP, shepherds genomic data through several steps to predict the 3-D dimensional structures of the proteins, and those predictions are graded based on their reliability. Statistically verifying the accuracy of the predictions is what sets the iGAP pipeline apart from others, according to Shindyalov. Scientists need to know the quality of the data that they will use in their research and the EOL data will be based on a measure of truth called a SCOP classification, considered the "gold standard" of protein prediction.

Having proved the value of the pipeline, Bourne and his colleagues began to consider the possibility of processing other genomes through it. This evolved into a large-scale project to annotate all 800 or so publicly available genomes, which, in mid-2002, Bourne named the Encyclopedia of Life.

Raw genetic sequence itself does not represent a daunting amount of data—the human genome contains some three billion base pairs, filling approximately three gigabytes, and some plant genomes are anticipated to be 10 times that size. However, processing genomes through the pipeline is not trivial. The actual annotation is an embarrassingly parallel computational problem that involves assigning on average a 10-kilobyte piece of biological data known as an Open Reading Frame, a translated DNA sequence potentially encoding a protein, to a processor. Each ORF produces approximately six megabytes of data. At most, processing 800 genomes will generate about 50 to 60 terabytes of data. That data would then be distilled into a single five-terabyte repository located at SDSC.

Bourne estimates that sending 800 genomes through the pipeline will be one of the largest computations ever attempted in biological science. To process all 800 would take a single processor working non-stop about 500 years—just for the first pass. Another way of looking at it is that each genome will average one day on 250 processors to complete. Then, since the resulting putative functional and structural annotation is based on comparison to experimental data, the amount of experimental data increases the genomes need to be recomputed. Fortunately for the EOL researchers, pipeline-type computation lends itself to parallel processing as well as distributed computing. The EOL is a charter application for the TeraGrid and was developed to make use of grid tools such as the AppLeS Parameter Sweep Template, developed in the Grid Research and Innovation Laboratory at SDSC .



DALLAS AND MARGARET HANNA © CALIFORNIA ACADEMY OF SCIENCES

EOL software has been adapted to a number of platforms, and is currently running at two sites, on six machines, including NPACI's Blue Horizon, in the United States and a cluster in Singapore. Research institutions in Japan, Korea, and China are scheduled to participate in the EOL in the near future. As of the end of February 2003, 71 genomes had been completely processed, and 80 others partially completed, organisms as diverse as anthrax (*Bacillus anthracis*) to the west nile virus (*Kunjin virus*) to frogs (*Xenopus laevis*), corn (*Zea mays*), and dogs (*Canis familiaris*).

"We expect the EOL to quickly gain momentum as more and more online information is added and quality scientific papers which have used the resource begin to appear," said Bourne. "By becoming a starting reference point and a gateway to information worldwide, we are not in any way challenging existing resources, we're actually making them more accessible."

—Cassie Ferguson

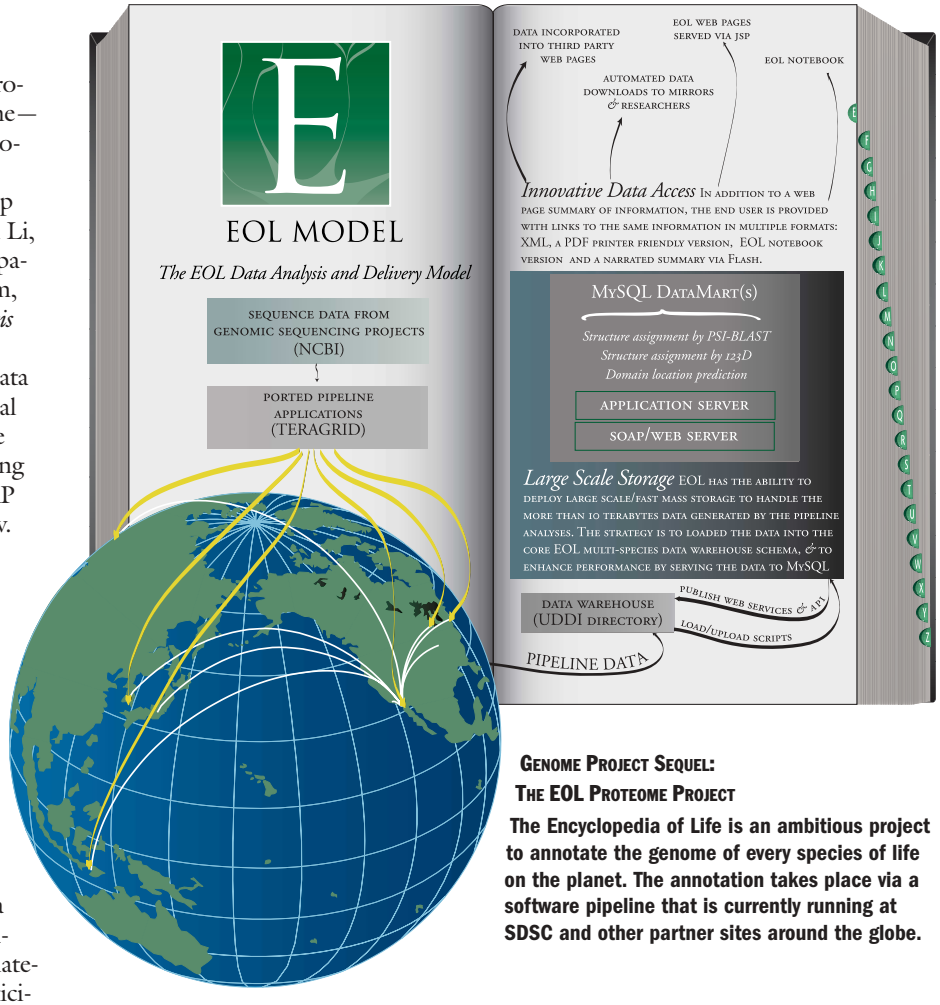


ILLUSTRATION BY JENNIFER MATTHEWS

GENOME PROJECT SEQUEL: THE EOL PROTEOME PROJECT

The Encyclopedia of Life is an ambitious project to annotate the genome of every species of life on the planet. The annotation takes place via a software pipeline that is currently running at SDSC and other partner sites around the globe.

Mining the Stars: The National Virtual Observatory

PROJECT LEADERS

ALEXANDER SZALAY
Johns Hopkins University

ROY WILLIAMS
Caltech

EXECUTIVE COMMITTEE

DAVID DEYOUNG
National Optical Astronomy Observatory

ROBERT HANISCH
Space Telescope Science Institute

GEORGE HELOU
Caltech/NASA Infrared Processing and Analysis Center

REAGAN MOORE
SDSC

ETHAN SCHREIER
Space Telescope Science Institute

The spectacular image on the cover of this issue of *EnVision* does not even begin to do justice to the information it contains. Its 144 million pixels would fill four IMAX theater screens at full resolution, which is one reason its creators and their collaborators at 17 institutions will be among the first users of the TeraGrid. "The National Virtual Observatory is a data-intensive, compute-intensive, and visualization-intensive project," says co-director Roy Williams, a Caltech computer scientist. "Our system design relies strongly on the data mining and grid technology that will be developed via the TeraGrid."

In little more than a year, the National Virtual Observatory (NVO) project has gained "even more momentum than we expected," says Williams, in the face of an ever-more-challenging avalanche of astronomical data. The project began in August 2001 under a five-year, $10 million Information Technology Research grant from the National Science Foundation (NSF). "The grant covers the framework for an NVO," says its other co-director, astronomer Alex Szalay of Johns Hopkins University, "and our objective is to do research about how such a thing could be made."

But the research is not taking place in a vacuum. "This is a matter of simultaneous implementation and planning in synergy with the astrophysical data community," Szalay says. The NVO collaboration involves 50 senior investigators from American institutions and has working liaisons with eight international observatory programs in Europe, Japan, and Australia. "We're co-leaders now of an International Virtual Observatory Alliance," says NVO Project Manager Robert Hanisch, who is also chairing that group.

THRIVING ON DATA

So the NVO idea has really taken off. But what is it? NVO is creating a huge virtual library of astronomical images, catalogs, measurements, and scientific publications, together with unified discovery and access services. Scientific exploration through computational methods is truly coming into its own—and real discoveries can be made under the paradigms of simulation and archive-based research, both of which are supported by NVO protocols.

Recent breakthroughs in ground- and space-based telescope, detector, and computer technology have enabled astronomers to undertake both wide and deep sky surveys in all wavelengths of the electromagnetic spectrum, from gamma rays to the far infrared. These are all producing multiple terabytes of data. A planned Large Synoptic Survey Telescope will produce 7-10 terabytes nightly and ultimately more than 10 petabytes annually. "That's why the National Academy of Sciences recommended the establishment of a National Virtual Observatory in 1999," Williams says. The idea was promoted by the NPACI Digital Sky project led by Caltech researcher (and NVO participant) Tom Prince. All of the major astronomical data archives in the nation are NVO participants.

To demonstrate the scientific possibilities NVO opens up, the project scientists have spent the first year producing several important data-federating and image-access tools, which they have incorporated into three "science prototype" demonstrations. "We thought of them as demonstrations only," Williams says, "but in fact one of them has already produced new astronomical knowledge."

GAMMA-RAY BURSTS, GALAXY SHAPES, AND BROWN DWARFS

The prototype projects were displayed in January at the American Astronomical Society annual meeting in Seattle. They are (1) a gamma-ray burst follow-up service, (2) a system for measurement and analysis of the variability of galaxy shapes in a cluster, and (3) a search for candidate "brown dwarf" stars—very faint, very small infrared stars that cannot sustain thermonuclear reactions, which may nevertheless carry a considerable proportion of the mass of the universe.

All three demonstrations relied on data collected by such sky surveys as the 2-Micron All-Sky Survey

MONTAGE PROJECT AT CACR, JPL, AND IPAC, CALTECH

WAR AND PEACE

Zooming in on the image on the cover of this magazine, this is a formation called the "War and Peace Nebula."

The Montage team is Tom Prince (Caltech; PI), Bruce Berriman (IPAC, Caltech; Project Manager), Anastasia Clower (IPAC, Caltech), John Good (IPAC, Caltech), Joe Jacob (JPL), Daniel S. Katz (JPL), and Roy Williams (CACR, Caltech). Post-production image processing was performed by Robert Hurt (IPAC, Caltech).

(2MASS), the Digital Palomar Observatory Sky Survey (DPOSS), and the Sloan Digital Sky Survey (SDSS). They will be further developed and presented, along with new prototypes, at the August 2003 General Assembly of the International Astronomical Union.

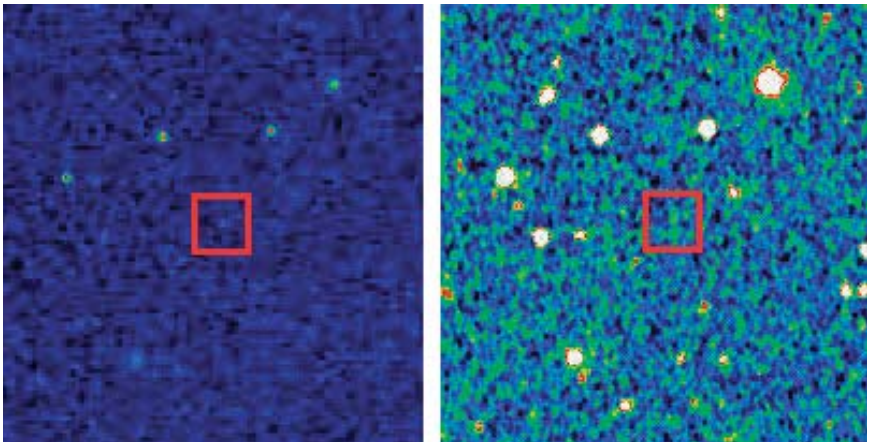
The gamma-ray burst follow-up service is intended to assemble data and alert observers whenever a gamma-ray burst occurs. These extremely energetic transient events occur all over the cosmos and are detected by orbiting gamma-ray observatories. They can also be seen in optical and lower wavelength "afterglows" if telescopes are pointed at them shortly after the gamma-ray signals are received. The service, which can also be used as a general tool to quickly access any interesting patch of sky, takes advantage of three protocols developed by NVO researchers: an interface called VOTable that quickly tabulates data from various surveys, a Cone Search protocol that queries databases about a specific sky location, and a Simple Image Access facility that quickly finds and combines images from the same part of the sky.

The galaxy morphology measurement and analysis service is also an interface to multiple surveys, and it will permit scientists to study the distribution of galaxy shapes (whirlpools, bars, spirals) and other parameters across galaxy clusters, which should enable understanding of galaxy and cluster evolution. The service involves combining multiple optical and X-ray data sets, computing galaxy shapes on the fly with sophisticated grid technology, and comparing them with optical and X-ray maps of the clusters.

The brown dwarf candidate search now uses two of the largest sky surveys, SDSS and 2MASS, and the first trial of the prototype on a small patch of sky came up with three previously unknown candidate stars fitting the "brown dwarf" profile. One has now been confirmed by subsequent observation, says Williams, which he finds especially pleasing. "We were only trying to demonstrate what kind of science might be done in the NVO environment, but this was like hitting the jackpot first time out."

NVO, MONTAGE, AND THE STORAGE RESOURCE BROKER

An example of a major use of TeraGrid technology in the NVO image domain is Montage, an astronomical mosaic service funded by the NASA Earth Science Technology Office Computational Technologies Project. Tom Prince of Caltech leads Montage development. Montage delivers mosaics of hundreds or thousands of smaller images from major sky surveys. These mosaics preserve the positional and photometric fidelity of the original images and can therefore be used directly in scientific analysis. The Montage team at the Caltech Center for Advanced Computational Research, the Jet Propulsion Laboratory, and the NASA Infrared Processing and Analysis Center (IPAC) at Caltech produced the cover image and the images on these pages. They worked with SDSC scientists George Kremenek and Leesa Brieger, who will also work with NVO TeraGrid projects.



MONTAGE PROJECT AT CACR, JPL, AND IPAC, CALTECH

BROWN DWARF

The brown dwarf discovered by the science prototype demonstration is seen here as a very faint source (in the center of the square area) in the 2MASS K band (2.2 μm) image. It is too faint to be seen in the Sloan Digital Sky Survey R band (red light).

Reagan Moore, an SDSC Distinguished Scientist on the NVO Executive Committee, says, "Montage and other NVO services are benefiting from work with the Storage Resource Broker developed in San Diego." For example, 10 terabytes of the 2MASS collection are available in HPSS at SDSC as well as in the HPSS archives at Caltech. "This assures data integrity and improves the reliability of data access by a factor of 10," Moore says, "because a request for data could be satisfied at either location." The SRB implements mechanisms for data aggregation in containers for bulk data movement, remote proxies for I/O command aggregations, bulk metadata manipulation, remote proxies for creating data subsets, and replication (location transparency) for data caching.

"This was particularly useful to us as we put the cover image mosaic together," said Bruce Berriman at IPAC, who directed the work. "Of course, not only SRB but also the IBM Blue Horizon machine was used," said Brieger, "where one mosaic can occupy 64 processors for three hours"

EFFICIENT WEB-BASED RESEARCH

"Our AAS demonstrations illustrated the way in which new science can result from the federation of very different archives," Williams concluded. "At present, NVO is developing standard protocols to increase the efficiency of this kind of research, but NVO will also make other web-based services available to researchers, including data quality control, online data and scientific publication access, and imaging modalities.

"It's also a sociological experiment," Williams said. "We are working out ways for large groups of researchers to collaborate with one another, and our approaches to the problems of data assimilation, ownership, credit, and quality are all dependent on a new level of cooperation that matches the promise of the science that can be done." —*Merry Maisel* ▼

Preserving Priceless Digital Holdings for the Library of Congress

PROJECT LEADERS

ARCOT RAJASEKAR, REAGAN MOORE
SDSC

MARTHA ANDERSON, JANE MANDELBAUM
Library of Congress

PARTICIPANTS

SHEAU YEN CHEN, CHARLIE COWART
SDSC

BAHA AKPINAR, HALLIE TRAVIS
Library of Congress

www.loc.gov
memory.loc.gov/ammem/amhome.html
www.npaci.edu/DICE/SRB
www.sdsc.edu/daks/index.html

The Library of Congress has assembled many important digital collections such as American Memory, a gateway to rich primary source materials ranging from rare photographs to historical documents including the Declaration of Independence. With rapidly-growing collections, the Library is working to build a repository to manage these digital holdings. Today, the powerful data grid technologies of the Storage Resource Broker (SRB), developed at the San Diego Supercomputer Center (SDSC) for scientific computing in the National Partnership for Advanced Computational Infrastructure (NPACI), are finding new uses beyond their original scientific applications as they help to preserve digital collections for the Library of Congress and other institutions.

The American Memory website offers more than 7.5 million digital records from more than 100 collections of books, manuscripts, films, maps, sound recordings, and photographs in the Library and other repositories. Items include encoded text and images as well as audio and video files varying in size from 25 kilobytes to five megabytes each, for a total of some eight terabytes of digital data.

SDSC and the Library are currently collaborating to evaluate the SDSC SRB data grid software for preservation and management of these priceless national digital collections—an irreplaceable part of the US national heritage—for decades and centuries into the future.

"We're entering an era in which digital libraries can be used to preserve a broad range of intellectual capital," said Reagan Moore, co-director of the Data and Knowledge Systems (DAKS) program at SDSC and leader of the NPACI Data-Intensive Computing Environments (DICE) thrust. "And beyond preservation, the ability to discover the information and knowledge content within digital holdings will add even greater value to these collections."

Martha Anderson of the Office of Strategic Initiatives at the Library of Congress comments, "The Library of Congress is collaborating with SDSC to explore emerging data grid technologies for preserving our digital collections. We're interested in how the SDSC SRB can be applied to the task of building a repository for managing Library of Congress digital holdings."

With the data grid technologies of the SDSC SRB, digital entities do not need to reside in the same physical location to be accessible and manageable by the Library, which is charged with the mission of preserving and providing access to its digital holdings over the long term. Data grid technologies can also help the Library develop the ability to preserve the integrity of its collections as the underlying storage technologies continue to evolve.

ABOUT THE LIBRARY OF CONGRESS

The mission of the Library of Congress is to sustain and preserve a universal collection of knowledge and creativity for future generations, and to make these resources available and useful to Congress and the American people. Founded in 1800, the Library of Congress is the world's largest library, with more than 124 million items in all formats on which information is recorded. The Library serves Congress and all Americans through 21 reading rooms on Capitol Hill and its popular website at **www.loc.gov.**

COLLECTION MANAGEMENT

The researchers will investigate the capabilities of the SDSC SRB to manage and to repurpose Library of Congress collections. Repurposing a collection involves giving users the ability to generate new views or perspectives of the digital holdings, rearranging or adding new metadata, or descriptive information about the collection, to create new collections. For

AMERICAN MEMORY COLLECTION, LIBRARY OF CONGRESS, PRINTS AND PHOTOGRAPHS DIVISION [PAN US GEOG-OREGON, NO. 30]

example, a user might want to gather the material in the American Memory collection that is relevant to, say, a landing on Mars. This material might involve NASA material on the mission and space vehicle, Congressional material on the budget debates involving the funding, and other material that puts the mission in historical context.

The collaboration includes the installation at the Library of Congress of the SDSC SRB software and the Metadata Catalog, which keeps track of each digital object. Library of Congress staff are building test collections and using them to evaluate the capabilities of the SDSC SRB data grid middleware to preserve both the collection and descriptive metadata; to enable a naming convention that spans the entire collection, no matter where its components are located; to merge different collections seamlessly into new virtual collections; and to control access. Library of Congress researchers are also interested in evaluating the ability of the SDSC SRB to interoperate with other systems using open standards.

"In addition to appraising the SDSC SRB in managing Library of Congress digital collections, we're looking forward to the research opportunities this collaboration will give us to understand how digital library, data grid, and persistent archive technologies can all be integrated in support of preservation of digital holdings," said Moore. "This will help extend our ability to preserve and explore intellectual capital." —*Paul Tooby*

AMERICAN MEMORY COLLECTION, LIBRARY OF CONGRESS, MANUSCRIPT DIVISION

PHOTOGRAPH ALBUM

Civil War photograph album, ca. 1861–65 (James Wadsworth Family Papers) showing *cartes de visite*, photographic calling cards very popular during the American Civil War. One of 7.5 million digital items in the American Memory collection of the Library of Congress. Library and SDSC researchers are evaluating the SDSC SRB for use in preserving these holdings over the long term.

PANORAMIC VIEW OF GREAT NORTHERN PACIFIC STEAMSHIP COMPANY'S TERMINALS, FLAVEL, OREGON, C1915

One of 7.5 million digital items in the American Memory collection at the Library of Congress. Library and SDSC researchers are evaluating the SDSC SRB for use in preserving these holdings over the long term.

Sapphire Worm Shatters Previous Speed Records for Spreading Through the Internet

PARTICIPANTS

David Moore
SDSC, UCSD

Vern Paxson
International Computer Science Institute; Lawrence Berkeley National Laboratory

Stefan Savage
UCSD

Colleen Shannon
SDSC

Stuart Staniford
Silicon Defense

Nicholas Weaver
Silicon Defense; UC Berkeley

A team of network security experts has determined that a computer worm dubbed "Sapphire" that attacked and hobbled the Internet in January was the fastest spreading computer worm ever recorded. In a technical paper released in February the experts report that the speed and nature of the Sapphire worm (also called Slammer) represent significant and worrisome milestones in the evolution of computer worms.

Computer scientists at San Diego Supercomputer Center (SDSC) at the University of Ccalifornia, San Diego; Silicon Defense; University of California, Berkeley; and the nonprofit International Computer Science Institute in Berkeley, found that the Sapphire worm doubled its numbers every 8.5 seconds during the explosive first minute of its attack. Within 10 minutes of debuting at 5:30 A.M. (UTC) on January 25, 2003 (9:30 P.M. PST, January 24), the worm was observed to have infected more than 75,000 vulnerable hosts. Thousands of other hosts may also have been infected worldwide. The infected hosts spewed billions of copies of the worm into cyberspace, significantly slowing Internet traffic and interfering with many business services that rely on the Internet.

"The Sapphire/Slammer worm represents a major new threat in computer worm technology, demonstrating that lightning-fast computer worms are not just a theoretical threat, but a reality," said Stuart Staniford, president and founder of Silicon Defense. "Although this particular computer worm did not carry a malicious payload, it did a lot of harm by spreading so aggressively and blocking networks."

The Sapphire worm's software instructions, at 376 bytes, are about the length of the text in this paragraph, or only one-tenth the size of the Code Red worm, which spread through the Internet in July 2001. Sapphire's tiny size enabled it to reproduce rapidly and also fit into a type of network "packet" that was sent one-way to potential victims, an aggressive approach designed to infect all vulnerable machines rapidly and saturate the Internet's bandwidth, the experts said. In comparison, the Code Red worm spread much more slowly, not only because it took longer to replicate but also because infected machines sent a different type of message to potential victims that required them to wait for responses before subsequently attacking other vulnerable machines.

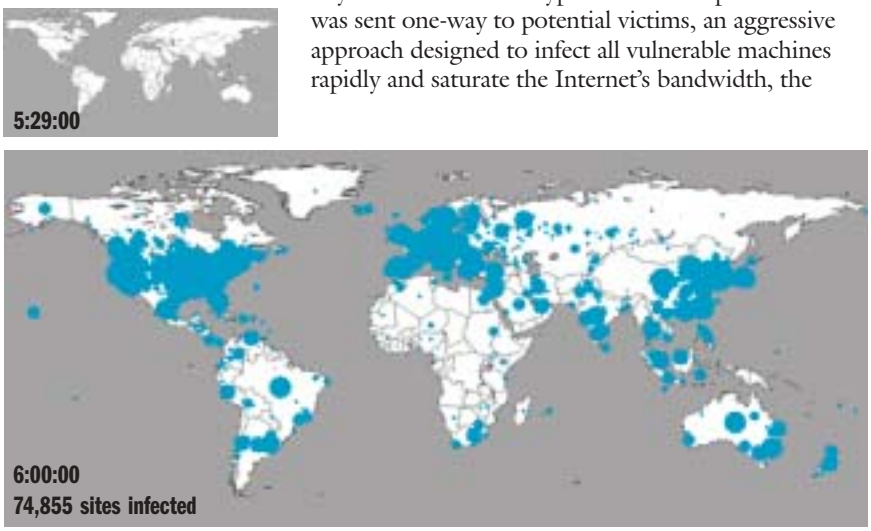
The Code Red worm ended up infecting 359,000 hosts, in contrast to the approximately 75,000 machines that Sapphire hit. However, Code Red took about 12 hours to do most of its dirty work, a snail's pace compared with the speedy Sapphire. The Code Red worm sent six copies of itself from each infected machine every second, in effect "scanning" the Internet randomly for vulnerable machines. In contrast, the speed with which the diminutive Sapphire worm copied itself and scanned the Internet for additional vulnerable hosts was limited only by the capacity of individual network connections.

"For example, the Sapphire worm infecting a computer with a one-megabit-per-second connection is capable of sending out 300 copies of itself each second," said Staniford. A single computer with a 100-megabit-per-second connection, found at many universities and large corporations, would allow the worm to scan 30,000 machines per second.

"The novel feature of this worm, compared to all the other worms we've studied, is its incredible speed—it flooded the Internet with copies of itself so aggressively that it basically clogged the available bandwidth and interfered with its own growth," said David Moore, an Internet researcher at SDSC's Cooperative Association for Internet Data Analysis (CAIDA) and graduate student at UCSD. "Although our colleagues at Silicon Defense and UC Berkeley had predicted the possibility of such high-speed worms on theoretical grounds, Sapphire is the first such incredibly fast worm to be released by computer hackers into the wild," said Moore.

Sapphire exploited a known vulnerability in Microsoft SQL servers used for database management, and MSDE 2000, a mini version of SQL for desktop use. Although Microsoft had made a patch available, many machines did not have the patch installed when Sapphire struck. Fortunately, even the successfully attacked machines were only temporarily out of service.

"Sapphire's greatest harm was caused by collateral damage—a denial of legitimate service by taking database servers out of operation and overloading networks," said Colleen Shannon, a CAIDA researcher. "At Sapphire's peak, it was scanning 55 million hosts per second, causing a computer version of freeway



(CAIDA/UCSD)

From Zero to 74,855 in Thirty-one Minutes

A map showing the number of infected Sapphire hosts in the period between 05:29 and 06:00 A.M. UTC immediately following the worm's debut on January 25, 2003. An animated version of the map is available at www.sdsc.edu/Press/03/020403_SAPPHIRE.html.

gridlock when all the available lanes are bumper-to-bumper." Many operators of infected computers shut down their machines, disconnected them from the Internet, installed the Microsoft patch, and turned them back on with few, if any, ill effects.

The team in California investigating the attack relied on data gathered by an array of Internet "telescopes" strategically placed at network junctions around the globe. These devices sampled billions of information-containing "packets," analogous to the way telescopes gather photons. With the Internet telescopes, the team found that nearly 43 percent of the machines that became infected are located in the United States, almost 12 percent are in South Korea, and more than 6 percent are in China.

Despite the worm's success in wreaking temporary havoc, the technical report analyzing Sapphire states that the worm's designers made several "mistakes" that significantly reduced the worm's distribution capability.

For example, the worm combined high-speed replication with a commonly used random number generator to send messages to every vulnerable server connected to the Internet. This so-called scanning behavior is much like a burglar randomly rattling doorknobs, looking for one that isn't locked. However, the authors made several mistakes in adapting the random number generator. Had there not been enough correct instructions to compensate for the mistakes, the errors would have prevented Sapphire from reaching large portions of the Internet.

The analysis of the worm revealed no intent to harm its infected hosts. "If the authors of Sapphire had desired, they could have made a slightly larger version that could have erased the hard drives of infected machines," said Nicholas Weaver, a researcher in the Computer Science Department at UC Berkeley. "Thankfully, that didn't occur." *—Rex Graham, Cara Sloman, Nadel Phelan, and Sarah Yang*

Researchers Find Unnecessary Traffic Saturating a Key Internet Root Server

Scientists at the San Diego Supercomputer Center (SDSC) analyzing traffic to one of the 13 Domain Name System (DNS) "root" servers at the heart of the Internet found that the server spends the majority of its time dealing with unnecessary queries. DNS root servers provide a critical link between users and the Internet's routing infrastructure by mapping text host names to numeric Internet Protocol (IP) addresses.

Researchers at the Cooperative Association for Internet Data Analysis (CAIDA) at SDSC conducted a detailed analysis of 152 million messages received on October 4, 2002, by a root server in California, and discovered that 98 percent of the queries it received during 24 hours were unnecessary. The researchers believe that the other 12 DNS root servers likely receive similarly large amounts of bad requests.

Some experts regard the system of 13 DNS root servers—the focus of several studies by CAIDA researchers—as a potential weak link in the global Internet. Spikes in DNS query traffic caused by distributed denial-of-service attacks are routinely handled by root server operators. Occasionally, as on October 21, 2002, all 13 root servers are attacked simultaneously.

Only about 2 percent of the 152 million queries received by the root server in California on October 4 were legitimate, while 98 percent were classified as unnecessary. CAIDA researchers are seeking to understand why any root server would receive such an enormous number of broken queries daily from lower level servers. "If the system were functioning properly, it seems that a single source should need to send no more than 1,000 or so queries to a root name server in a 24 hour period," said CAIDA researcher Duane Wessels. "Yet we see millions of broken queries from certain sources."

Wessels categorized all the queries received by the California root server on October 4, 2002, into nine types. About 70 percent of all the queries were either identical, or repeat requests for addresses within the same domain. It is as if a telephone user were dialing directory assistance to get the phone numbers of certain businesses, and repeating the directory assistance calls again and again. Lower level servers and Internet service providers (ISPs) could save—or cache—these responses from root servers, improving overall Domain Name Service performance.

About 12 percent of the queries received by the root server on October 4 were for nonexistent top-level domains, such as "**.elvis**", "**.corp**", and "**.localhost**". Registered top-level domains include country codes such as "**.au**" for Australia, "**.jp**" for Japan, or "**.us**" for the United States, as well as generic domains such as "**.com**", "**.net**", and "**.edu**". In addition, 7 percent of all the queries already contained an IP address instead of a host name, which made the job of mapping it to an IP address unnecessary.

Researchers believe that many bad requests occur because organizations have misconfigured packet filters and firewalls, security mechanisms intended to restrict certain types of network traffic. When packet filters and firewalls allow outgoing DNS queries, but block the resulting incoming responses, software on the inside of the firewall can make the same DNS queries over and over, waiting for responses that can't get through. Name server operators can use new tools such as dnstop, a software program written by Wessels and available from CAIDA, which detects and warns of these and other misconfigurations, significantly reducing bad requests. *—Rex Graham*

Building a 'Memory' for the National Science Digital Library

PROJECT LEADER
DAVID FULKER
NSDL

PARTICIPANTS
CHARLES COWART, AMARNATH GUPTA, MEVLUT KURUL, REAGAN MOORE, ARCOT RAJASEKAR
SDSC

nsdl.org

www.sdsc.edu/daks

The National Science Digital Library (NSDL) opened its virtual doors in December of 2002. Whether users are experts or novices, the NSDL can help a third-grade teacher make a quick Internet stop for targeted materials, or help others take thought-provoking strolls through rich science resources, eliminating barriers to accessing educational resources while transmitting the excitement of scientific discovery to educators and learners alike. SDSC is a partner in developing the NSDL's technical infrastructure, providing a stable "memory" for the Library through long-term archiving services, which make use of new extensions to the SDSC Storage Resource Broker (SRB) data grid software. Data and Knowledge Systems (DAKS) researchers are also providing knowledge-based services that will allow educators better navigation and more precise searches as well as more advanced features including the ability to compare multiple curricula.

Sponsored by the National Science Foundation (NSF), the NSDL is expected to grow into the world's largest digital library of science, technology, engineering, and mathematics (STEM) resources and services for education. "The NSDL is the nation's most comprehensive effort to support science education through a digital library," said Reagan Moore, co-director of SDSC's DAKS program and leader of the NPACI Data-Intensive Computing Environments (DICE) thrust. "Along with the work of more than 100 partners, SDSC and NPACI technologies are helping develop the NSDL as a structured education layer over the Web, a significant step in the use of data grid and knowledge-based technologies to support education."

The "virtual library" of the NSDL—a sort of "library of libraries"—will serve students and teachers at all levels of education, from pre-K-12, undergraduate, and graduate to life-long learning. "The NSDL provides an organized, coherent view of a wide range of interactive learning materials, including data, visualization tools, and other rich resources," said David Fulker, NSDL director. "Offering both breadth and depth of inquiry, the Library will help users discover specific educational materials as well as enhanced methods of teaching."

NSDL SEARCH SCREEN

Results for a search for science education resources on the topic of "fossil fish." Note the "Archived Version" button, which links educators to SRB-archived versions of NSDL resources, providing reliable long-term access to resources they have integrated into curricula, even if the original changes, moves, or disappears from the Web.

ACCESSING SCIENCE EDUCATION RESOURCES

Far more than a static collection of online information in a website, the NSDL offers visitors a portal—an entire Web system of services, content, and applications—that enables a wide range of users to participate in activities from targeted searches and simple browsing to collaboration in curriculum creation and discovery of new teaching methods and uses for NSDL resources. And behind the simple NSDL interface is a growing array of powerful, integrated services that, transparently to users, connect them to a wealth of distributed educational resources.

At the heart of the Library is a powerful search service that helps users efficiently discover resources for STEM education. Suppose a K–12 educator is seeking information on fossil fish to use in the classroom. Using the NSDL search engine, she is directed to a number of curricular resources on fossil fish through links to educational resources that may consist of text, video, and multimedia, or even entire websites and other digital libraries.

When our educator initiates her NSDL search for resources on fossil fish, the portal uses the search engine developed by the University of Massachusetts at Amherst, which searches both text and the metadata repository at Cornell University—the electronic "card catalog" for the Library—describing the relevant Library holdings. The educator sees URLs that link her to Web-based educational resources, which may be distributed at various participating sites.

Even in this basic search, a number of services are

involved that leverage the strengths of a number of different organizations. In addition to the above partners, the NSDL Core Integration Team includes the University Corporation for Atmospheric Research and Columbia University, with contributions from Eastern Michigan University, the University of California, Santa Barbara, the University of Colorado at Boulder, Syracuse University, and other partners.

BUILDING A PERSISTENT NSDL ARCHIVE

An interesting aspect of this large-scale collaboration is the speed with which the services are evolving. The initial plan was for SDSC simply to archive the metadata from the Cornell repository. "But then we realized we could add the feature of retrieving the actual archived curriculum module, and we demonstrated this as an end-to-end service," said DAKS researcher Charles Cowart. "Given a Library record, we can provide back through the Web not only identifying links or metadata but also an archived copy of the full digital object."

When users search for resources, there is an "Archived Version," button next to records for which an archived copy is available. Clicking this button accesses the SDSC service, which then locates and retrieves the archived curriculum module. To make the service work, a Web crawler has been incorporated to archive not just the metadata but the educational resources themselves into an SRB archive, which will be maintained at SDSC. "What is making this possible is the maturity of the SRB," said Cowart. "Building on this reliable middleware makes it practical to add this end-to-end retrieval capability."

Now suppose our educator has previously developed a curriculum that incorporates the fossil fish information she found through the NSDL, which is provided by the Paleontological Research Institution. "One of the big frustrations educators face with the Web is that access is unreliable," said Fulker. "After they've spent a lot of time finding things on the Web and integrating them into curricula, too often they find that the Web resources have changed or disappeared." To realize the potential of the NSDL, educators need long-term access to resources—a "memory" for the NSDL as its dynamic Web-based resources evolve. Through archived copies, SDSC's NSDL archive will enable the K–12 educator to have sustained access to the particular version her curriculum relies on. "SDSC has achieved a remarkably quick implementation of the initial preservation model, which will store periodic snapshots of Library resources that users will be able to retrieve as needed," said Fulker. The initial implementation provides access to the latest version, and the researchers are working with NSDL colleagues to design ways to specify and access prior versions.

SDSC persistent archiving technologies are part of an architecture that will maintain the authenticity of the archived NSDL collection, while minimizing the effort needed to incorporate new technology as hardware and software continue to evolve. Additional uses of persistent archiving in the NSDL will include disaster recovery for NSDL metadata, validation of NSDL repository metadata, a curricula re-creation service, addressing intellectual property and access issues, and a knowledge research repository. "Eventually, we would like to make the NSDL persistent archive available for knowledge mining activities on the NSF TeraGrid, which is designed to support the analysis of 10 terabytes of data per hour on a 4 teraflops compute engine," said Moore.

ONLINE EDUCATIONAL RESOURCE

An educational resource on fossil fish teeth from sharks and other common fish that is available online as part of the NSDL from a museum collection at the Paleontological Research Institution.

NOT YOUR MOTHER'S CARD CATALOG

The initial NSDL focus is to offer important exemplars that extend the concepts of what a library can be. While libraries have always offered an environment that enhances creativity, with skilled librarians providing reserve shelves and resources especially prepared for certain uses, libraries have also sometimes been seen as "passive" places.

"The NSDL can help invigorate traditional library practices to produce user-driven, highly 'local' views, or sub-collections, that are rooted in the huge resources of the total NSDL," said Fulker. The dimensions along which such views can be created include age or education level, discipline, or various media formats. "Another important view we hope to develop, with the aid of appropriate partners, is a mapping of NSDL resources onto national and state standards for science education," said Fulker. This can even be implemented at the district or teacher level, providing a scalable "education standards view" of the NSDL.

In extending library functions, the NSDL is seeking to improve access, making it easier to wade through the overwhelming amount of information to find desired items. This puts the emphasis on organization and the capability for the NSDL to provide different views of the collections.

The traditional approach to this is the card catalog, which contains information about the library holdings such as title, author, year, and so on, organized in alphabetical order. In the digital world, this tool has been greatly extended into rich electronic catalogs that are more flexibly searchable. The contents of a digital "card catalog" are called metadata—descriptive information about the digital library entities—which can extend far beyond basic title-author-year information.

In the NSDL, various approaches to handling searches are being explored. One way of organizing and relating the library contents is to develop metadata across all the disciplines. While this method will enable broad searches, it has limitations in specialized searches. For example, suppose an educator wants to search the "map room" being developed in the NSDL. Because the electronic "card catalog" system or metadata standard adopted for all NSDL collections—the widely used Dublin core metadata standard—does not support geo-referenced metadata, in order to be able to find maps and other geo-referenced objects, the NSDL will have to extend the metadata standard to include this information. Similarly, for other disciplines not presently included in the Dublin core metadata standard, the researchers will have to find ways to specify the relevant information.

Other search approaches avoid metadata entirely and rely on inferential or other search mechanisms. It is possible that future methods will rely on both metadata as well as inferential searches. And beyond text-based information, the library researchers are developing ways to extend search capabilities to include scientific data sets or useful computer programs as objects of interest.

In addition to persistent archiving services, to help improve library searches, DAKS researchers Amarnath Gupta and Mevlut Kurul are participating in development of ontology services, which make use of a controlled, hierarchical vocabulary for describing a knowledge system—on top of the basic NSDL infrastructure. The ontologies are being developed by connecting NSDL holdings to related educational concepts. The system will be able to support basic services such as an ontology-based search to select audience-appropriate items covering a given topic, or to evaluate the quality of instructional material such as a textbook with respect to a given topic. In addition, the system will support more complex services such as comparison of multiple curricula for students at a specified grade level.

NASA HUMAN SPACE FLIGHT

EARTH SCIENCE IMAGE IN THE NSDL

Oblique view of the Sinai Peninsula from Egypt (left) to Saudi Arabia (right) captured by astronauts aboard the Space Shuttle Columbia while servicing the Hubble telescope in March of 2002. Note darker Precambrian metamorphic rock on the southern end of the peninsula contrasting with lighter sedimentary rock to the north and more recent sands and gravels along the coasts.

"We have demonstrated a process that begins with a representation of a portion of the American Association for the Advancement of Science (AAAS) Project 2061 concept space, from which we extracted keywords from a learning topic on cells and applied them against an index of the contents of related, linked NSDL documents," said Gupta, director of the DAKS Advanced Query Processing Lab. The AAAS Project 2061—an important guideline for science education—is a long-term initiative to reform K-12 science, mathematics, and technology education nationwide, and the project has produced the *AAAS Atlas of Science Literacy*.

The initial demonstration of the concept space was done with a set of concepts from a page of this *Atlas of Science Literacy*, and a search index of some 150 URLs of related NSDL articles that can explain the concepts. Clicking on an atlas topic returns a list of relevant documents (in the eXtensible Markup Language, or XML,

so they can be further queried to refine the search). To move toward full production capability in the NSDL knowledge management system, the researchers will undertake the major task of incorporating the complete AAAS Project 2061 concept space, and archiving all of the digital entities in the NSDL, in collaboration with the more than 100 projects developing collections or other Library services. The index will be generated from the documents in the SDSC-developed NSDL persistent archive.

Current research issues include extending the system capabilities to traverse the links between related concepts in the atlas, including grade-level classification of the materials in the search results, and expanding or narrowing the scope of the search. This involves questions of which and how many concept keywords to use, and how many NSDL learning units to query across for a given user search. The researchers will vary the granularity of each, both expanding the number of learning units from which keywords are taken, and going deeper in the index to look for matches. "We're trying to learn what balance of breadth in keywords and learning topics and what depth of query will give the most relevant search results," said Gupta.

ENHANCING EDUCATION ON A LARGE SCALE

The overall NSDL project is organized around four pillars: content, including a set of exemplary collections; innovation, inspiring and incorporating advances in both content and digital library research; education, including a network of educators so that improvements in content and digital library technology can evolve hand-in-hand with advances in pedagogy; and finally, partnership, extending the NSDL environment to include publishers, professional societies, and others during this time of dynamic, technology-driven change in education and libraries.

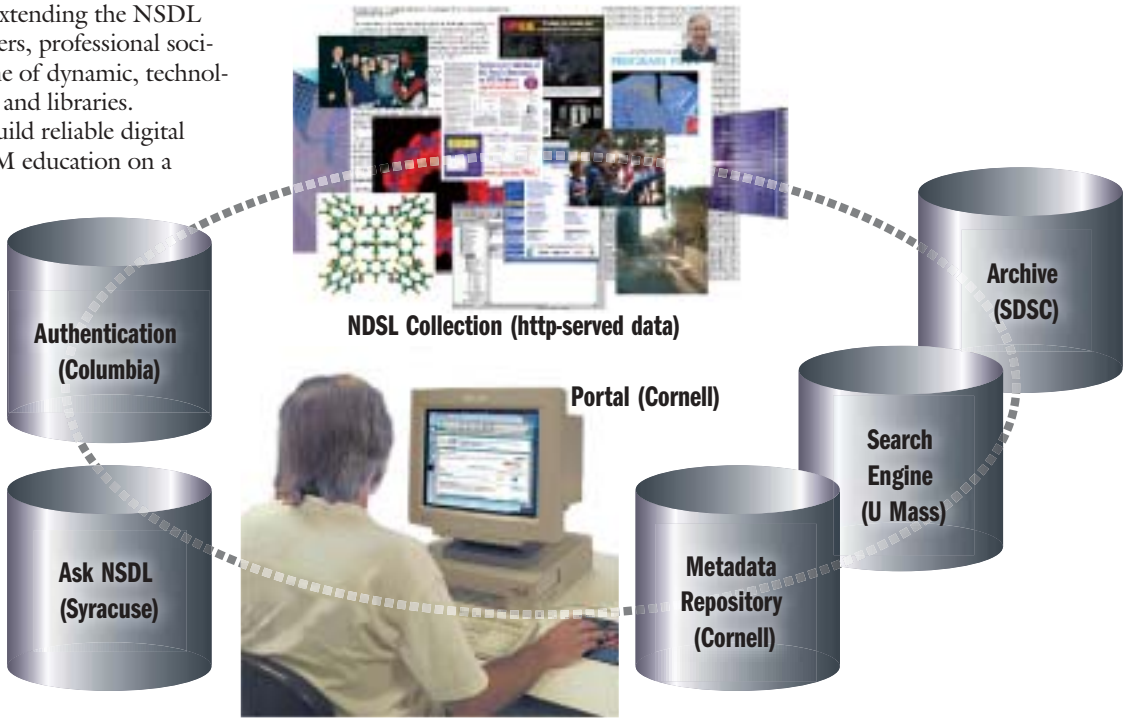
"The NSDL is working to build reliable digital library services to enhance STEM education on a very large scale," said Fulker. Because the NSDL will encompass all of science, technology, engineering, and mathematics education, this enormous scope will have great impact on education.

"Hand-in-hand with these practical services, the NSDL offers an environment of innovation to stimulate and incorporate advances in digital library research," said Fulker. "But we don't think the Library will work in a 'build it and they will come' way. We're actively working to engage educators and help them create valuable resources that will give them a sense of ownership." A central NSDL purpose is forming networks of educators who can make the NSDL a stimulus and tool to extend and enhance education. "It's clear that constructivist approaches where people participate are key," said Fulker. "You reach larger numbers of students and awaken greater excitement and interest so that they taste the creative side of this human endeavor—learning science by doing science."

The underlying technology that supports the NSDL Search Service is designed to scale up to accommodate a very large volume of data, eventually matching many thousands of users with millions of catalog records. As the technology progresses, the NSDL will offer a wider range of views of the holdings matched to a given grade level, as well as views tailored for educators, librarians, the public, and others, amplifying the Library's ability to meet user needs in ways that have not yet been imagined.

"SDSC has provided demonstrations of important archiving capabilities for the NSDL," said Moore. "The NSDL is a work in progress, and we look forward to collaborating in the large task of bringing these capabilities into full-scale production."

The NSDL will continue to transform as the technologies and education evolve in tandem. "We invite others beyond current NSDL partners to participate by contributing new educational resources," said Fulker. "As the NSDL grows, we hope it will become ever more desirable to participate." —*Paul Tooby* ▼



NSDL; PHOTO ILLUSTRATION BY GAIL W. BAMBER, SDSC

NSDL Core Infrastructure Architecture

Researchers are building core integration services for the NSDL to support online collections of education material. The digital library will manage descriptive metadata about material that is provided via the Web. Services include browsing on the collection, search across the digital entities stored at registered URLs, and a persistent archive of the registered material. When access is not available to the original material, the most recent version is provided from the archive.

ONLINE NEWS ABOUT NPACI AND SDSC

www.npaci.edu/Online

World's Fastest Network to Connect TeraGrid Sites

Fiber optic links between Los Angeles and Chicago have been "lit up" to form the cross-country network backbone for the National Science Foundation's $88 million TeraGrid project. Technicians are sending the first test data packets racing across the network, which boasts an unprecedented bandwidth—orders of magnitude faster than a typical dial-up Internet connection and four times faster than existing research networks.

At 40 gigabits per second, the new "backplane," developed in partnership with Qwest Communications, will connect the resources of the TeraGrid, a multiyear effort to build and deploy the world's largest, fastest, distributed computing infrastructure for open scientific research. Scientists will use the TeraGrid to make fundamental discoveries in fields as varied as biomedicine, global climate, and astrophysics. The first applications will begin to use the TeraGrid capabilities from all sites this spring. When completed, the TeraGrid will include 20 teraflops (trillion floating point operations per second) of computing power, facilities capable of managing and storing nearly one petabyte (one quadrillion bytes) of data, high-resolution visualization environments, and toolkits for grid computing. *(v 7.5)*

Version 2.0 of the Storage Resource Broker (SRB) Data Management Middleware Released

SDSC has released version 2.0.0 of the popular SDSC Storage Resource Broker middleware package, which enables scientists to create, manage, and collaborate with unified "virtual data collections" that encompass heterogeneous data resources distributed over a network. While existing ways of doing things are preserved for current users, major enhancements "under the hood" give version 2.0 a large number of faster and more powerful services. SDSC SRB version 2.0.0 along with the user manual and release notes are available online.

Interest is growing in the SDSC SRB software because of the need to integrate, manage, and access explosively growing data collections in many fields. There are currently more than 200 registered users of the SDSC SRB at more than 50 sites who share the common need to manage, integrate, and collaborate with large data sets. *(v 7.4)*

TACC to Upgrade Its IBM Power4 System: Teraflops Performance and Large Shared Memory Capability

The Texas Advanced Computing Center (TACC) on the campus of the University of Texas at Austin, will enhance the capacity of its IBM Power4 system, known as "Longhorn." After the upgrades are complete, Longhorn will have 224 Power4 processors with a theoretical peak performance of 1.16 teraflops. The total memory of the system will be 0.5 terabyes, with 5 terabytes total disk in the high-speed parallel I/O file system. TACC Director Jay Boisseau said, "The new system will be especially interesting because it has different nodes for achieving maximum performance on both shared and distributed memory applications, and at both small and large numbers of processors." *(v 7.3)*

SDSC Selected as a National Internet2 Technology Evaluation Center

SDSC is one of three sites selected by Internet2 as a national Internet2 Technology Evaluation Center (ITEC). The mission of the center will be to test and evaluate leading-edge technologies for high-performance Internet2 networks—working with developers to test and refine network hardware and software for optimal end-to-end network performance up to 10 gigabits per second. Internet2 is a consortium led by more than 200 U.S. universities, working with industry and government to develop and deploy advanced Internet applications and technologies. *(v 7.4)*

Leadership Happenings: Changes

Alan Blatecky, a national and international leader in grid computing and networking, will join SDSC as executive director in the spring of 2003. Blatecky currently directs the National Science Foundation's Middleware Initiative, a pioneering program that is developing the foundation for the next generation of information technology and cyberinfrastructure. Blatecky also co-directs the inter-agency MAGIC (Middleware and Grid Infrastructure Coordination) program group, which coordinates middleware and grid technologies throughout the U.S. government and coordinates with international efforts.

Andrew Chien of UCSD was recently named to SDSC and NPACI Director Fran Berman's team of Strategic Advisors. He joins half a dozen distinguished scientists who are helping to provide regular guidance on various topics of interest to both SDSC and NPACI. Chien has just returned to full-time faculty status at UCSD from a two-year leave during which he founded Entropia, Inc., a leading grid software company focusing on desktop grids.

Nancy Marlin replaced Pieter Frick as the NPACI Institutional Oversight Board (IOB) representative from San Diego State University. The Institutional Oversight Board, chaired by UCSD Chancellor Robert Dynes, consists of individuals from institutions with major cost-sharing commitments to NPACI or that reflect the diversity and strength of the partnership. The IOB's chief responsibilities are to ensure that commitments from individual institutions are kept and to appoint the External Visiting Committee.

A second change was made by the IOB regarding the chairmanship of the External Visiting Committee (EVC). Dona Crawford of Lawrence Livermore National Laboratory will now chair the EVC, replacing Ann Hayes of Los Alamos National Laboratory, who has retired from membership after five years of service. Wayne Pfeiffer is stepping down from the NPACI Executive Committee in order to focus on his research efforts at SDSC, and William R. Martin of the University of Michigan, presently on the Executive Committee, is joining the NPACI Leadership Team. *(v 6.24, 7.1, 7.2)*

Leadership Happenings: Honors

The British Computer Society will present the 2002 Lovelace Medal to Carl Kesselman, NPACI's chief software architect and director of the University of Southern California's Information Sciences Institute's Center for Grid Technologies, in May 2003 in London. Kesselman will receive the medal along with Ian Foster of the University of Chicago and Argonne National Laboratory's Mathematics and Computer Science Division for their joint work in leading the Globus

To view the full Online article, append the issue number to the URL: www.npaci.edu/online/vX.X

Project and grid computing. The Lovelace Medal is presented to individuals who have made contributions of major significance in the advancement of information systems. Previous recipients of the medal include Doug Engelbart, developer of the computer mouse and computer windows; and Linus Torvalds, developer of the Linux operating system.

NPACI and SDSC Director Fran Berman has been appointed a member of the California Council on Science and Technology. Berman was appointed to a three-year term on the council, a not-for-profit corporation established by the California State Assembly to examine urgent public policy questions relating to science and technology in California, including the state's competitiveness.

Former SDSC and NPACI Director Sid Karin, now senior advisor to SDSC and a professor of computer science and engineering at UCSD, has been elected to Fellowship in the Association for Computing Machinery, the scientific and professional society for computer science and information technology. *(v 6.24, 7.1, 7.2)*

SDSC's Data and Knowledge Systems Program Launches Advanced Database Projects Lab

As scientific data sets have grown more massive, scientists often find that their traditional tools bog down or don't work—it can simply take too long to access the subset of data the researchers need. To solve this problem, the DAKS program at SDSC, a world leader in large-scale scientific data management and knowledge discovery, has recently organized an Advanced Database Projects Lab.

Chaitan Baru, co-director of the Data and Knowledge Systems program said, "The Advanced Database Projects lab will provide an important component of DAKS research and support scientific research from the TeraGrid to large-scale collaborations like the Biomedical Informatics Research Network, the Encyclopedia of Life, and the National Virtual Observatory, and small but extremely valuable databases like the Great Apes Phenome Project." *(v 7.3)*

TACC and Platform Computing Collaborate to Develop 'Grid of Grids' for Scientific Research

Platform Computing Inc. and the Texas Advanced Computing Center (TACC) announced that they will collaborate on research and development of next-generation software technologies for grid computing. This partnership will focus deployment and use of scientific applications and experiments on grids. The University of Texas at Austin is the largest university in the U.S., and TACC is building a university-wide UT Grid that will connect the myriad of clusters, workstations, visualization systems, and storage devices of researchers and departments to TACC's high-end facilities. Platform LSF and Platform MultiCluster are viewed as key technologies for enabling researchers to share HPC resources and to execute codes that span multiple HPC systems. *(v 6.25)*

NPACI Demonstrates Success at SC2002

NPACI demonstrated its computational infrastructure and scientific successes to a record gathering of high-performance computing enthusiasts at SC2002 in Baltimore. Researchers participated both in the technical portion of the conference and showcased their work in the NPACI booth.

The SDSC and NPACI booth, which took five days to construct and a day to disassemble, covered a forty-by-forty-foot area, and featured four demo areas and colorful graphics related to NPACI science successes. Behind the scenes, a support team from SDSC assembled and maintained the booth's computational infrastructure, which included a diversity of hardware and software used for demonstrations designed to demand the most of the technology.

NPACI researchers Henrique Andrade and Alan Sussman of the University of Maryland, and Joel Saltz and Tahsin Kurc of Ohio State University were awarded SC2002 Best Student Technical Paper with "Active Proxy-G: Optimizing the Query Execution Process in the Grid." *(v 6.24)*

APRIL 2003

6-8 Passive and Active Measurement Workshop (PAM2003), hosted by NLANR/MNA, San Diego Supercomputer Center, La Jolla, CA

10-11 AAAS Colloquium on Science and Technology Policy, Washington, DC

22-26 17th Annual International Parallel and Distributed Processing Symposium (IPDPS 2003), Nice, FRANCE

30-May 2 Alliance All-Hands meeting, Urbana, IL

MAY

12-15 Third IEEE /ACM International Symposium on Cluster Computing and the Grid (CCGrid 2003), Tokyo, Japan

18-21 International Conference on Computational Science and Its Applications (ICCSA 2003), Montreal, Canada

JUNE

2-4 The International Conference on Computational Science 2003 (ICCS 2003), Melbourne, Australia

22-24 HPDC-12 (held in conjunction with GGF-8), Seattle, WA

22-25 Global Grid Forum (held in conjunction with GGF-8), Seattle, WA

23-25 Global Gird Forum (GGF8), Seattle, WA

24-27 International Supercomputer Conference (ISC2003), Heidelburg, Germany

29-July 3 Intelligent Systems for Molecular Biology (ISMB), Brisbane, Australia

For more information and events, see the SDSC calendar on the Web:

www.sdsc.edu/Calendar

BACK COVER

NPACI PARTNERS

University of California, San Diego/
San Diego Supercomputer Center
California Institute of Technology
University of Texas, Austin
University of Michigan
University of California, Berkeley
University of California, Santa Barbara
University of Southern California/Information
Sciences Institute
University of Virginia
Baylor College of Medicine
California State University/San Diego State
University
University of California, Davis
University of California, Irvine
University of California, Los Angeles
The Johns Hopkins University
University of Maryland
Montana State University
University of New Mexico/Long-Term Ecological
Research Network
New York University
Ohio State University
Oregon State University
Rice University
Rutgers, The State University of New Jersey
Salk Institute for Biological Studies
The Scripps Research Institute
Stanford University
University of Tennessee
Washington University
University of Wisconsin, Madison
Center for Advanced Research in Biotechnology
Jet Propulsion Laboratory
Kitt Peak National Observatory
Lawrence Berkeley National Laboratory/National
Energy Research Scientific Computing Center
Lawrence Livermore National Laboratory
Los Alamos National Laboratory
University of Massachusetts
Pacific Northwest National
Laboratory/Environmental Molecular Sciences
Laboratory
University of Pennsylvania
BioComputing Unit, Centro Nacional de
Biotecnología, Madrid, Spain
Parallel Computing Center, Royal Institute of
Technology, Stockholm, Sweden
University of Queensland, Brisbane, Queensland,
Australia
Research Institute for the Management of Archives
and Libraries, University of Urbino, Italy

KEIKO NOMURA, JAMES ROTTMAN, HIDEAKI TSUTSUI, AND MIKE BAILEY, UCSD

SHORT-WAVE INSTABILITY OF A VORTEX PAIR

The 3-D instabilities of a pair of counter-rotating vortices were investigated using computer simulations at SDSC. Such flows may be encountered in the wakes of aircraft, where they can be hazardous to other aircraft. Researchers are studying how these vortices break down and decay. The visualization shows a short-wave instability in which the vortices deform sinusoidally and antisymmetrically with wavelength on the order of the initial vortex spacing. Transverse vortex structures form between the two main vortices leading to a rapid transition to turbulence. The image was produced by professors Keiko Nomura and James Rottman and graduate student researcher Hideaki Tsutsui, Department of Mechanical and Aerospace Engineering, UCSD. 3-D volume rendering was performed by Mike Bailey of SDSC, adjunct professor of Mechanical Engineering and Computer Science at UCSD. ▼

SUBSCRIPTIONS

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

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

eradmin@sdsc.edu, 858-534-5111

NAME

TITLE

INSTITUTION

ADDRESS

CITY, STATE, AND ZIP

COUNTRY

E-MAIL

NPACI NATIONAL PARTNERSHIP FOR ADVANCED COMPUTATIONAL INFRASTRUCTURE

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

ADDRESS SERVICE REQUESTED

PRESORTED STANDARD
U.S. POSTAGE
PAID
WESTERN GRAPHICS