CRD report

News from the LBNL Computational Research Division — June 2004

Anatomy of a Web(bed) Legend: Virtual Frog @ 10

In the wild, a frog may live to 10 years, assuming it survives tadpolehood and doesn't get eaten by a bird or a fish or some other creature. On the Web, though, a virtual frog named "Fluffy" has easily notched its tenth year despite millions of dissections. Launched in June1994, Berkeley Lab's Virtual Frog Dissection Kit Web site allows users to virtually dissect a frog without all that smelly formaldehyde of high school science class. Since then, more than 15 million people in more than 130 countries have visited the frog at <http://dsd.lbl.gov/Frog/>.

"We know it's still being used because whenever the server goes down, we get a lot of email from around the world, mainly from teachers who use the site in their classes," said Bill Johnston, currently head of the DOE's Energy Sciences Network (ESnet).

Johnston and other members of the former Imaging Technologies Group, now part of the Lab's Distributed Systems Department, developed the Virtual Frog. While the Virtual Frog has remained virtually unchanged over the past decade, the research behind the site has advanced by leaps and bounds, leading to Berkeley Lab's prominent role in the development of Grid technologies.

"The Virtual Frog was a direct result of our early work in wide-area distributed computing and visualization," Johnston said. "And this has led directly to our involvement in developing collaborative technologies and Grid applications."

The frog was originally spawned by LBNL's work in developing software to create visual renderings of sets of diagnostic cross-section images used in medicine, allowing scientists to create 3D renderings from MRI data. Once

(continued on page 2)

Scientific Data Management Center Helping Scientists Focus on Science, Not Data

While terascale supercomputers are giving computational scientists unparalleled research capabilities, these systems are also producing huge amounts of data to be managed. Similar situations confront researchers using massive experimental facilities, where new experiments will be generating unprecedented quantities of scientific data. As a result, researchers often spend more time trying to find ways to manage their data instead of analyzing them.

To help scientists make effective and efficient use of these facilities and the resulting data, the Scientific Data Management Group in Berkeley Lab's Computational Research Division is leading DOE's project to coordinate the development and deployment of scientific data management software. Arie Shoshani, leader of the LBNL group, is also the lead principle investigator for the Scientific Data Management Center (SDM) funded under DOE's SciDAC program. The center consists of four DOE laboratories (ANL, LBNL, LLNL, ORNL) and four universities (GTech, NCSU, NWU, SDSC).

"Our purpose is not only to achieve efficient storage and access to the data, but also to enhance the effective use of the scientist's time by eliminating unproductive simulations, by providing specialized data-mining techniques, by streamlining time-consuming tasks, and by automating the scientist's workflows," Shoshani wrote in a report on the project's achievements to date. "Our approach is to provide an integrated scientific data management framework where components can be chosen by the scientists and applied to their specific domains. By overcoming the data management bottlenecks and unnecessary information-technology overhead through the use of this integrated framework, scientists are freed to concentrate on their science and achieve new scientific insights."

Scientific research typically takes place in two phases: data collection/generation and data analysis. In the collection/generation phase, large datasets are generated by simulation programs or collected from experiments. This requires efficient parallel data systems that can keep up with the volumes of data generated. In the analysis phase, efficient indexes and effective analysis tools are necessary to find and focus on the information that can be extracted from the data, and the knowledge learned from that information.

Being able to analyze the data as they are generated is also important. For example, a scientist running a thousand-time-step 3D simulation can benefit from analyzing the data

(continued on page 2)

Raquel Romano Named Alvarez Fellow in Computational Science

As Bahram Parvin (left) and Juan Meza (right) look on, Horst Simon congratulates Raquel Romano on her selection as the Alaverez Fellow in Computatonal Science.

Raquel Romano, a member of the Imaging and Informatics Group in the Computational Research Division, has been selected as the next Luis W. Alvarez Post-Doctoral Fellow in Computational Science. Here she is congratulated by Horst Simon, associate lab director for Computing Sciences. After earning her bachelor's degree in mathematics from Harvard, Romano earned both her master's and Ph.D. in computer science at MIT. At Berkeley Lab, she is working to develop computer vision and image processing tools to help biologists in the Life Sciences Division better understand why some cells become cancerous. Read more about Romano at <http://www.lbl.gov/CS/html/romano.html>.

What is CRD Report?

This is CRD Report, a publication highlighting recent achievements by staff members in Berkeley Lab's Computational Research Division. Distributed every other month via email and posted on the Web at http://crd.lbl.gov/DOEresources, CRD Report may be freely distributed. CRD Report is edited by Jon Bashor, JBashor@lbl.gov or 510-486-5849.

Scientific Data Management

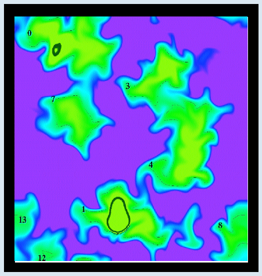
(continued from p.1)

from individual steps in order to steer the simulation, saving unnecessary computation and accelerating scientific discovery. This requires sophisticated workflow tools, as well as efficient dataflow capabilities to move large volumes of data between the analysis components. For these reasons, the team uses an integrated framework that provides a scientific workflow capability, supports data mining and analysis tools, and accelerates storage access and data searching.

Progress at LBNL

Since the SDM Center was launched three years ago, the team has adopted, improved and applied various data management technologies to several scientific application areas, concentrating on typical scenarios provided by scientists from different disciplines. Not only did the team learn the important aspects of the data management problems from the scientist's point of view, but also provided solutions that led to actual results. In addition to overall direction and coordination of the center, the LBNL team has applied advanced indexing techniques and storage management to several application domains

• A new specialized method for indexing high-dimensional data was applied to mesh data using bitmaps and achieved more than a ten-fold speedup in generating regions and tracking them over time. The figure here shows the regions identified by their assigned numbers for a combustion application. The regions are tracked over time using the bitmaps to efficiently generate a movie of flame front progression.



The key to this achievement is that this method works just as efficiently for selection conditions over multiple measures, a problem previously unsolved with any known indexing techniques. This bitmap-based indexing method was also applied to find collisions (events) of interest in high energy physics applications, and is currently used in the STAR experiment at BNL. This capability reduced the amount of data scientists have to sift through, reducing search times from weeks to hours.

• A new software module called Storage Resource Manager (SRM) has been used to provide wide-area access to the HPSS mass storage system. This enables scientific programs to remotely stream files from HPSS into the program's disk space without explicitly requesting them, greatly simplifying the scientist's task when using very large datasets. SRMs are also used in high energy physics and climate applications to move data robustly from data-generating sites to data-consuming sites. This task of moving terabytes reliably was prone to error and required constant monitoring and recovery from errors. The SRM technology now automates both the transfer and the recovery from transient errors, saving the scientist many hours of wasted time.

"The bottom line is that it is over 100 times faster than what I was doing before," said Michael Wehner, a climate researcher at LBNL and SRM "power user." "More importantly. I can see a path that will essentially reduce my own time spent on file transfers to zero in the development of the climate model database."

Contact Arie Shoshani at shoshani@lbl.gov.

Virtual Frog (continued from p.1)

the software was developed, Johnston said, he began to think about what to do with it. The line of thinking led to DOE's education programs, and then to secondary schools. "Obviously, the choice was to create a frog."

The first step was to obtain a specimen, purchased at a Berkeley pet shop and then euthanized following a strict, UC-approved protocol for handling animals for experiments. Once the amphibian was dead, lab assistant Katie Brennan brought it out to the team with her hands clasped around the creature. Opening them slowly, she declared, "Voila – Fluffy!" A star was named.

Next, the challenge was to use LBNL's research MRI machine to create the layered images. "No matter what they did, the images came out mushy," Johnston said. The cause turned out to be the dark green spots on the frog's back. Containing iron oxide pigment, the spots set up magnetic fields that wreaked havoc on the imaging system. So, a cryotome machine with a swinging diamond blade was used to make 130 extremely thin slices of tissue. After each slice was removed, the remaining portion of the frog was photographed and digitized. The images were then segmentized by using a pen to outline critical organs or systems on each of the 130 slides.

"That is what allows you to manipulate 11 different anatomical systems," said David Robertson, who developed the Virtual Frog Dissection Kit that makes the whole thing work. "This was one of the first, if not the first, 3D rendering applications on the Web."

Because many people were still using 9,600 baud modems, Robertson compressed the images to make them more easily accessible. With no search engines and minimal marketing, the Virtual Frog leapt to prominence by word of mouth. "It rapidly shot up to 1.5 million hits a year," Robertson added. "And it continues to get up to 40,000 hits a day."

David Robertson shows how the compression files used to create the Virtual Frog make the Web site accessible using a smart phone.

In a case of what is old becoming new again, the compression techniques that allowed quick display of images over a 9600 baud modem now enable people with "smart phones" to call up the Virtual Frog using the browsers on the phones, which typically have low-bandwidth wireless connections.

"I don't think any of us realized how popular – and how enduring the site would become," Johnston said. "What started as something of an experiment has become an institution."

DISCLAIMER
This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California. Ernest Orlando Lawrence Berkeley National Laboratory is an equal opportunity employer.