CRD report

News from the LBNL Computational Research Division

March 2005

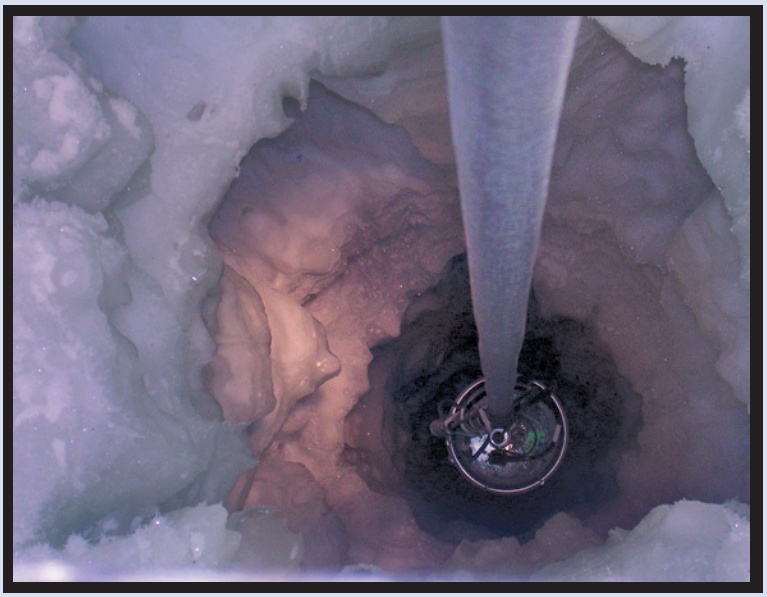
LBNL Team Takes Software to the End of the Earth

Developing robust, reliable data acquisition software for high-energy physics experiments is always a challenge, but developing such software for an experiment expected to run for up to 15 years while buried in the Antarctic ice poses unique problems.

But a team led by Chuck McParland of CRD's Distributed Scientific Tools Group has risen to the occasion. The first kilometer-long string of 60 detectors recently buried near the South Pole is already recording light pulses as the experiment seeks out evidence of neutrinos.

The experiment, an international undertaking by 26 institutes, is known as IceCube as the strings of detectors will cover a cubic kilometer of ice. The detectors, which will number 4,800 once installation is completed in 4-5 years, will serve as a telescope

(continued on page 2)



The first string of detectors containing data acquisition software developed at LBNL is lowered into the Antarctic ice as the IceCube experiment begins its search for neutrinos.

CRD's Juan Meza Named to National Research Council Panel

Juan Meza, head of the High Performance Computing Research Department in CRD, has been appointed as a member of the National Research Council Panel I: Modeling and Simulation, Systems Engineering, and Cost and Risk Analysis, one of 10 panels to review NASA's Capability Roadmaps. The panel will be chaired by Dr. Nancy Leveson, professor of aeronautics and astronautics at the Massachusetts Institute of Technology.

Juan Meza

The National Research Council is part of the National Academies of Sciences. The Research Council was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of further knowledge and advising the federal government.

Biological Data Management and Technology Center Marks First Year

The Biological Data Management and Technology Center (BDMTC) at Lawrence Berkeley National Laboratory marked its first anniversary with the release of the Integrated Microbial Genomes (IMG) system, a complex biological data management system BDMTC developed in collaboration with the Microbial Genome Analysis Program (MGAP) at the Joint Genome Institute (JGI).

Victor Markowitz

As a community resource, IMG integrates JGI's microbial genome data with publicly available microbial genome data, providing a powerful comparative context for microbial genome analysis. At JGI, an enhanced version of IMG provides support for advanced data curation and annotation carried out by MGAP scientists.

While IMG is the first academic "product" BDMTC undertook, its success "demonstrates the viability of the center's rationale," said BDMTC head Victor Markowitz, who launched the center in January 2004. "BDMTC is based on the premise that addressing effectively biological data management challenges requires extensive data management and system development experience and expertise consolidated in a central core," according to Markowitz.

Particularly encouraging has been the response of MGAP scientists to the systematic approach followed in developing IMG, from gathering and analysis of user requirements through the public release of the system.

"Although I was closely involved for years in the development of another microbial genome data system at Integrated Genomics, this is the first time I have experienced a well organized process in which requirements are documented, clarified, and continuously refined, and development follows a strict yet clear and predictable schedule" said Nikos Kyrpides, head of MGAP and IMG's scientific lead. "I fully appreciate the value of a disciplined development process which I now consider as critical to ensuring that the system addresses the needs of the scientific users."

In presentations about IMG, Kyrpides emphasizes the benefits biologists gain from system documentation, the development process,

(continued on page 3)

LBNL Team Takes Software to the End of the Earth

(continued from page 1)

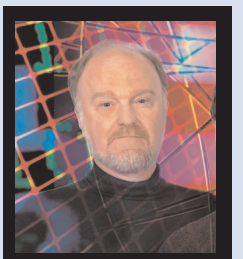
designed to study the high-energy variety of the ghostlike subatomic particles known as neutrinos. Originating from the Milky Way and beyond and traveling to Earth virtually unobstructed, these high-energy neutrinos serve as windows back through time and should provide new insight into questions about the nature of dark matter, the origin of cosmic rays, and other cosmic issues.

> "It's a very interesting computer science problem in that we have a distributed group of people working to put together a good software design with high quality code."

Berkeley Lab researchers were responsible for the unique electronics package inside the digital optical modules (DOMs) that will enable IceCube to pick out the rare signal of a high-energy neutrino colliding with a molecule of water. A DOM is a pressurized glass sphere the size of a basketball that houses an optical sensor, called a photomultiplier tube, which can detect photons and convert them into electronic signals that scientists can analyze.

Equipped with onboard control, processing and communications hardware and software, and connected in long strings of 60 each via an electrical cable, the DOMs can detect neutrinos with energies ranging from 200 billion to one quadrillion or more electron volts. In January and February, the first IceCube cable, with its 60 DOMs, was lowered down into a hole drilled through the Antarctic ice using jets of hot water. Plans call for a total of 80 strings of DOMS to be put in place over the next five years. The Antarctic summer season, during which the weather is "mild" enough for work on IceCube to proceed, lasts only from mid-October to mid-February. After that, winter sets in and the climate is much too harsh for any outdoor work.

While the operating environment offers a unique challenge, McParland said, "The biggest challenge is bridging the gap between the hardware side of the project and the software side." The hardware developers needed some of the data acquisition software to test their systems, but that was just one aspect of the overall software project. "Getting the testing and the actual data acquisition components to all work together is a challenging job," he said.



Chuck McParland

Adding to the difficulty was the push to engineer software that would support the 10- to 15-year lifespan of the experiment. "This means we had to put a lot more effort into the design of the software and use more modern programming practices than are typically used for other experiments of this sort," McParland said. Most of the acquisition software is being designed by an LBNL-led team which includes Chris Day, Simon Patton, Akbar Mokhtarani, Artur Muratus, Keith Beattie, Martin Stoufer, Arthur Jones, David Hays and John Jacobsen of Berkeley Lab, as well as researchers at the University of Wisconsin, Madison; Penn State University and the University of Delaware.

The software is being developed in two pieces. The first piece goes "down the hole," loaded on each of the DOMs. The system is so designed that the software in each module can be updated and, thanks to programmable logic, the circuit boards can also be reprogrammed to change the behavior of the detectors.

"Since the first string was installed we are already seeing events – cones of light that lit up all of the modules," McParland said.

At the top of each detector string will be a PC, where the second piece of software will be installed. Together, the PCs will form a pyramid of processors, with each performing initial filtering of the data before sending them to a satellite for transmission to what will eventually be a 120-processor farm.

The flow of data is in the shape of waveforms describing each event. Each waveform is time stamped. This allows them to be compared, with accuracy to within five nanoseconds, so that researchers can see how and when a particle passed near each detector. Calibrating the system to allow this comparison required the team to develop a new protocol to stop all communication between the modules – a task beyond the capability of off-the-shelf protocols, McParland said.

Once the data have been collected and an interesting event found, scientists will actually trigger the experiment after the fact, McParland said. This is achieved by breaking apart the collected data and going back through it to find the waveforms depicting the event.

Another problem faced by the developers was dealing with impurities in the glass used in the spheres and photomultiplier tubes. As the impurities decay, they give off light pulses, which meant that the software had to filter out 50 to 100 times more incidental signals than actual scientific data. "There's just lots of background stuff you have to get rid of," McParland said.

IceCube will field about 20 times the detector power of its predecessor, another South Pole high-energy neutrino telescope called AMANDA (Antarctic Muon And Neutrino Detector Array). McParland was also part of the AMANDA project, and spent five weeks at the South Pole as part of that effort.

While his group's role in IceCube will begin to wind down in 2006, for the time being the work is both very demanding and proceeding very well, McParland said.

"Overall it's a very interesting computer science problem in that we have a distributed group of people working to put together a good software design with high quality code," he noted.

CRD Report

CRD Report is published every other month, highlighting recent achievements by staff members in Berkeley Lab's Computational Research Division. Distributed 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.

Biological Data Management and Technology Center Marks First Year

(continued from page 1)

and even data modeling abstractions, a view he hopes more biologists will start to share.

Rationale for BDMTC

Biological data management involves data generation and acquisition, data modeling, data integration and data analysis. Data management poses challenges on several fronts. First, there are the increasing amounts of experimental data generated by life science applications. Next is the difficulty of qualifying data generated using inherently imprecise tools and techniques. Finally, there is the complexity of integrating data residing in diverse and poorly correlated repositories.

At research institutions such as LBNL and the University of California, San Francisco (UCSF), biological data management systems have typically been developed with an eye toward rapid development and low cost. This often meant that minimal consideration was given to requirements analysis, system development practices, system evolution, maintenance and scalability. While such as approach was perceived as less expensive because it could be achieved without experienced data management professionals and software engineers, the savings came at the expense of overall system quality, including reliability, maintenance and evolution.

The problems associated with academic systems and software have been recognized and addressed in NIH reports, in particular "The Biomedical Information Science and Technology Initiative" (BISTI) report prepared by the Working Group on Biomedical Computing Advisory Committee to the NIH Director and the NIH Roadmap for Accelerating Medical Discovery to Improve Health. Both documents recommend employing advanced data management technologies for developing interoperable biomedical databases and software engineering practices for delivering robust and reliable systems and tools.

Following NIH's recommendations requires expertise in several areas, such as data modeling, data integration, database administration, data sharing and security, software engineering, and software and data management quality control. Due to the complexity and cost involved, few public institutions can afford to acquire such expertise. Therefore a central core such as BDMTC could provide an effective solution to this problem. BDMTC's premise is also consistent with DOE's Genome to Life (GTL) program, which envisions consolidated computing infrastructure facilities in the form of software, biocomputing and data centers. In particular, a "seamless and effectively centralized capability to deal with data" in the form of data centers collecting and effectively integrating large scale biological data is seen as key to GTL's success.

Exploring Partnership Possibilities

Over the course of its first year, members of BDMTC approached a number of academic organizations in the Bay Area, both to assess their data management needs and to identify potential areas for collaboration. The organizations included the Berkeley Structural Genomics Center (BSGC), the Joint Genome Institute (JGI), the P50 Integrative Cancer Biology Program (ICBP) in the Life Sciences Division at LBNL, and the Immune Tolerance Network (ITN) at UCSF.

The analysis of JGI's data management goals subsequently led to the development of IMG. BSGC's data management needs, in particular in the area of experimental data tracking and work scheduling, were examined in order to prepare the Laboratory Information Management Systems (LIMS) and data management component of BSGC's PSI-II application for a Large Scale Structural Genomics Center. ICBP's data management core provided a concrete framework for exploring caBIG related opportunities for providing better data management support for NCI sponsored programs and centers.

BDMTC has also pursued collaborations with the Immune Tolerance Network (ITN) at UCSF and was part of UC Berkeley's proposal for a National Center for Biomedical Computing (NCBC): the former was not finalized because of budget cuts, and the latter was not selected for funding. However these initiatives provide additional evidence for BDMTC's potential to establish collaborations. In particular, NIH's call for establishing NCBC envisions software development and data management cores similar in scope to BDMTC.

Challenges and Plans

While there is clearly a growing need for enhanced data management tools, there is also a preference among many academic life science groups for a "do it yourself" approach, rather than collaborating with other groups or centers. An additional challenge is posed by the emphasis put on experimental results over data management, which may entail reconciling the relatively low budgets these groups assign to data management and the cost associated with outside collaborations.

In 2005, BDMTC will continue to pursue collaboration opportunities, primarily at LBNL, UC Berkeley and UCSF. Helping research groups realize that collaborations could lead to potentially higher quality data management results, reduced effort duplication, and savings coming from sharing resources and expertise will be part of BDMTC's outreach efforts.

Developing productive collaborations will require a change in the way groups tend to operate, Markowitz said. Life science groups would benefit from a higher level of collaboration in the area of biological data management system and bioinformatics tool development. Data management and software engineering groups need to improve their ability to support life science applications through enhanced understanding of these applications. Prerequisites for such an endeavor include finding incentives to encourage collaborations and raising the awareness of the critical role played by biological data management in competing for large-scale projects or centers such as those envisioned by the GTL and NIH programs.

Center's Microbial Data Management System to Be Featured at Conference

The Integrated Microbial Genomes (IMG) System developed by CRD's Biological Data Management and Technology Center (BDMTC) will be a featured demonstration at the 13th annual international conference on Intelligent Systems for Molecular Biology. The conference, sponsored by the International Society for Computational Biology, will be held June 25-29 in Detroit.