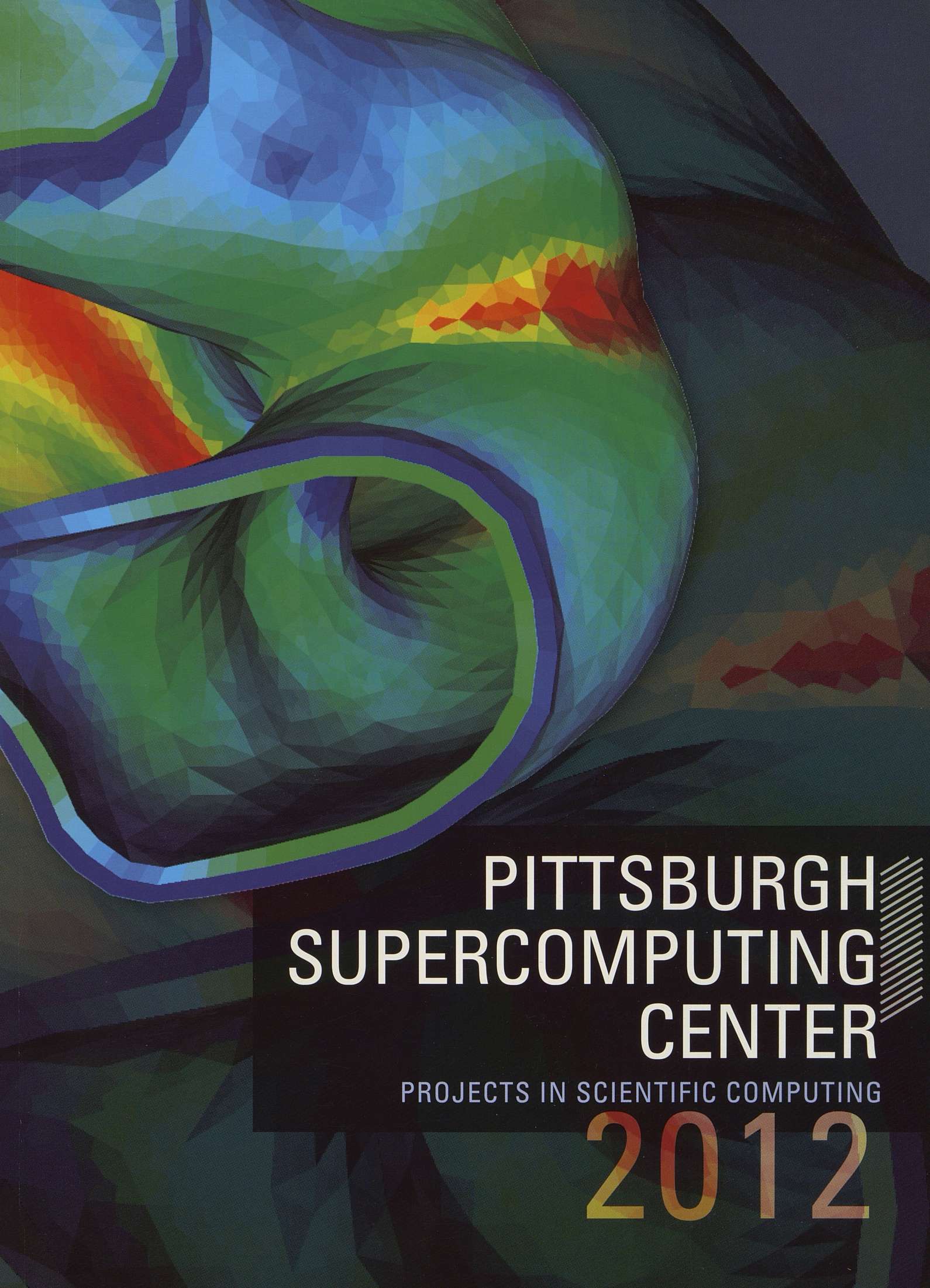
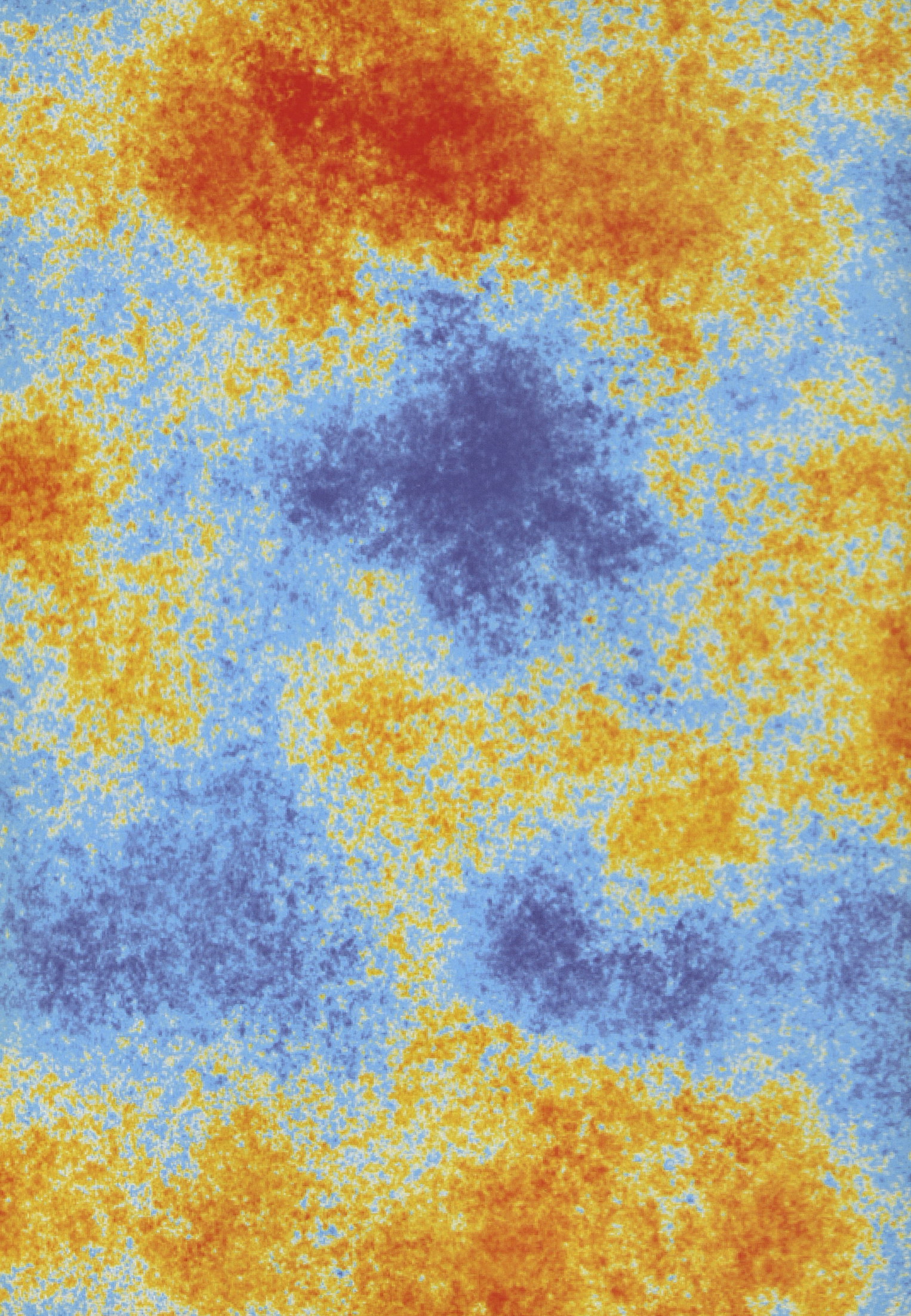


# PITTSBURGH SUPERCOMPUTING CENTER

## PROJECTS IN SCIENTIFIC COMPUTING

# 2012



# PSC.EDU/12

The Pittsburgh Supercomputing Center provides university, government and industrial researchers with access to several of the most powerful systems for high-performance computing, communications and data storage and handling available to scientists and engineers nationwide for unclassified research. PSC advances the state-of-the-art in high-performance computing, communications and informatics and offers a flexible environment for solving the largest and most challenging problems in computational science. As a leading partner in XSEDE, the Extreme Science and Engineering Discovery Environment, the National Science Foundation's cyberinfrastructure program, PSC works with other XSEDE participants to harness the full range of information technologies to enable discovery in U.S. science and engineering.

www.psc.edu
412.268.4960

PITTSBURGH SUPERCOMPUTING CENTER

# **FOREWORD** FROM THE DIRECTORS

**Ralph Roskies** (left) and **Michael Levine,** PSC co-scientific directors.

We're glad once again to present some of the year's accomplishments at the Pittsburgh Supercomputing Center (PSC). It's been gratifying to see our center's vision for the importance of shared memory bear fruit, as Blacklight – the world's largest shared-memory system (p. 4), now in its second year – has proven its value across a range of fields. We're pleased also to highlight our innovative disk-based, data-handling system, the Data Supercell (p. 4), and a new system, Sherlock (p. 6), specialized for graph analytics.

Our biomedical program is extended with new funding (p. 10), which includes a $1.1 million two-year extension for our Anton program. This collaboration with D. E. Shaw Research has already produced remarkable new insights in protein function (pp. 26-31).

In genomics, Blacklight is helping to open the possibilities of next-generation sequence data in new areas of research – behavioral genomics and metagenomics (pp. 18-21) – that didn't exist a few years ago. Through XSEDE's Extended Collaborative Support Services program, PSC scientist Phil Blood has provided nearly every software tool used in assembly and analysis in pre-compiled form, enhancing the advantages of Blacklight's large shared memory for this work.

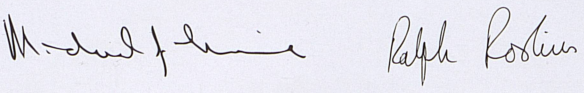
Computer trading on Wall Street – and the concerns it raises for U.S. financial systems – has been grabbing headlines, with media interest heightened by this year's Facebook IPO difficulties. Large-scale data analysis by Mao Ye and colleagues (pp. 22-25) yielded findings that illuminate these concerns. This work – enabled by Blacklight and Gordon at the San Diego Supercomputer Center – has caught the attention of policy-makers, including testimony before a U.S. Senate subcommittee.

The Los Angeles region has taken a lead among U.S. metropolitan areas in planning to face the social changes accruing with climate change. With Blacklight handling a large part of the computing, Alex Hall and his colleagues (pp. 32-35) produced a first-of-its-kind study that quantifies the local effects of global climate change, laying the groundwork for regional planning to help the community adapt to these changes.

Blacklight also has proven to be powerful in cosmological modeling, with very large-scale work investigating the Universe's phase change from "the Dark Ages" to reionization (pp. 40-43), and in quantum chemistry (pp. 36-39), with modeling of "conjugated polymers" that hold potential to revolutionize semiconductors.

PSC continues to be a resource for research and education in Pennsylvania (p. 7). Our networking group (pp. 12-13) serves the Pennsylvania-West Virginia region and carries out nationally recognized research in next-generation Internet resources. Along with advancing technology, we also help to educate the upcoming generation of scientists and science-literate citizens (pp. 8-9).

Our staff – second-to-none in talent and experience in high-performance computing – makes all this possible. We're grateful for support from the National Science Foundation, the U.S. Department of Energy, the National Institutes of Health, the Commonwealth of Pennsylvania and many others.



# CONTENTS

# BLACKLIGHT & THE DATA SUPERCELL

**The world's largest shared-memory supercomputer, PSC's Blacklight, has helped to open XSEDE resources to many non-traditional HPC projects**

Times are changing for high-performance computing (HPC) research, as fields of study that haven't traditionally used HPC have begun taking advantage of these powerful tools. This is especially true for PSC's Blacklight, an SGI® Altix® UV1000 system acquired in July 2010, with help from a $2.8 million award from the National Science Foundation. As the largest shared-memory system in the world, Blacklight has opened new capability for U.S. scientists and engineers.

"Blacklight has opened new doors to high-performance computation," said PSC scientific directors Michael Levine and Ralph Roskies, "and rapidly become a force across a wide and interesting spectrum of fields."

This was part of the plan for the NSF's XSEDE (Extreme Science and Engineering Discovery Environment) program, which launched in July 2011. The program this year took large steps toward this objective, with a number of non-traditional projects – the common denominator being the need to process and analyze large amounts of data – using XSEDE resources, especially Blacklight, to arrive at new insights.

Among these, described in this booklet, are work that analyzes huge quantities of finance-trading data to arrive at important new findings concerning non-beneficial effects of computer trading of stocks (see pp. 22-25). Several projects in assembly and analysis of "next-generation" sequence data have found that Blacklight's shared memory is uniquely well suited to advance work in this field (see pp. 18-21).

Shared memory offers a large advantage for many data-intensive applications because all of the system's memory can be directly accessed from all of its processors, as opposed to distributed memory (in which each processor's memory is directly accessed only by that processor). Because all processors share a single view of data, a shared-memory system is, relatively speaking, easy to program and use.

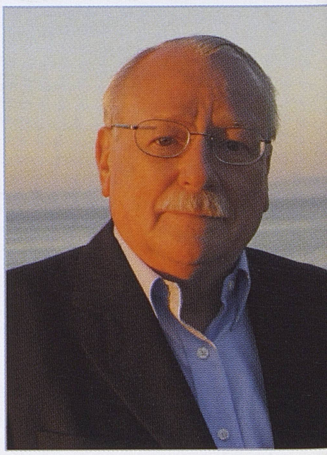
## The Data Supercell

PSC this year deployed a disk-based file repository and data-management system, the Data Supercell (DSC). This innovative technology provides significant advantages over tape-based archiving. The PSC development team – Paul Nowoczynski, Jared Yanovich, Zhihui Zhang, Jason Sommerfield, J. Ray Scott, and Michael Levine – exploited increasing cost-effectiveness of commodity disk technologies, and adapted sophisticated PSC file-system software (called SLASH2) to use with DSC. A patent application is under review.

"The Data Supercell is a unique technology, building on the cost-effectiveness of disk and the capabilities of PSC's SLASH2 file system," said Michael Levine and Ralph Roskies, PSC scientific directors. "It enables more efficient, flexible analyses of very large-scale datasets."

Intended especially to serve users of large scientific datasets, such as many XSEDE researchers, the DSC's initial capacity, four petabytes, can be expanded as needed. In comparison with tape-based archiving, DSC facilitates very fast data transfer (latency 10,000 times less than and bandwidth many times more than tape), while it also incorporates high reliability and security.

Departments at the University of Pittsburgh, Carnegie Mellon and Drexel are now using DSC, and researchers with large genomic datasets, produced through Galaxy, a web-based platform for bioinformatics at Penn State, are currently using 470 terabytes of DSC storage.



# Creating National Cyberinfrastructure

## As a leading partner in XSEDE, the most powerful collection of integrated digital resources and services in the world, PSC helps to shape the vision and progress of U.S. science and engineering

**PSC's directors** (l to r), who oversee day-to-day PSC operations and help to coordinate PSC's role in XSEDE: **Nick Nystrom**, director, strategic applications; **Sergiu Sanielevici**, director, scientific applications & user support; **Bob Stock**, PSC associate director; **David Kapcin**, director of financial affairs; **Wendy Huntoon**, director of networking; **David Moses**, executive director; **J. Ray Scott**, director, systems & operations. (Not pictured here, **Cheryl Begandy**, director of education, outreach and training.)

Through XSEDE, the Extreme Science and Engineering Discovery Environment, the NSF cyberinfrastructure program that launched in July 2011, PSC extends its active role in the development of national cyberinfrastructure. PSC scientific co-director Ralph Roskies is a co-principal investigator of XSEDE and co-leads its Extended Collaborative Support Services (ECSS). "ECSS staff work both with user groups in fields familiar with high-performance computing," says Roskies "and with the XSEDE outreach team to reach user groups, communities and digital services that are new to HPC."

Other PSC staff lead many areas of the comprehensive XSEDE program. Janet Brown, who manages PSC's network research, leads the XSEDE Systems and Software Engineering team that oversees the software environment that integrates resources among many providers. As manager of XSEDE Outreach Services, PSC manager of education, outreach and training Laura McGinnis leads programs that help to prepare the next generation of computational scientists.

PSC's security officer, Jim Marsteller, is the Incident Response Lead for XSEDE. Wendy Huntoon, PSC director of networking, is XSEDE networking liaison for the software development and iteration office. Ken Hackworth, PSC's user relations coordinator, leads the XSEDE allocations process by which research proposals are reviewed and evaluated to receive grants of computational time on XSEDE resources. PSC scientist Sergiu Sanielevici, director of scientific applications and user support for PSC, leads the Novel and Innovative Projects area of XSEDE's ECSS effort, which focuses on development of projects in fields or from institutions and communities that can exploit advanced computing but haven't traditionally used it.

Jim Kasdorf, who joined PSC scientific directors Michael Levine and Ralph Roskies in writing the proposal that established PSC, is PSC's director of special projects, involved in planning and coordination of many PSC initiatives.

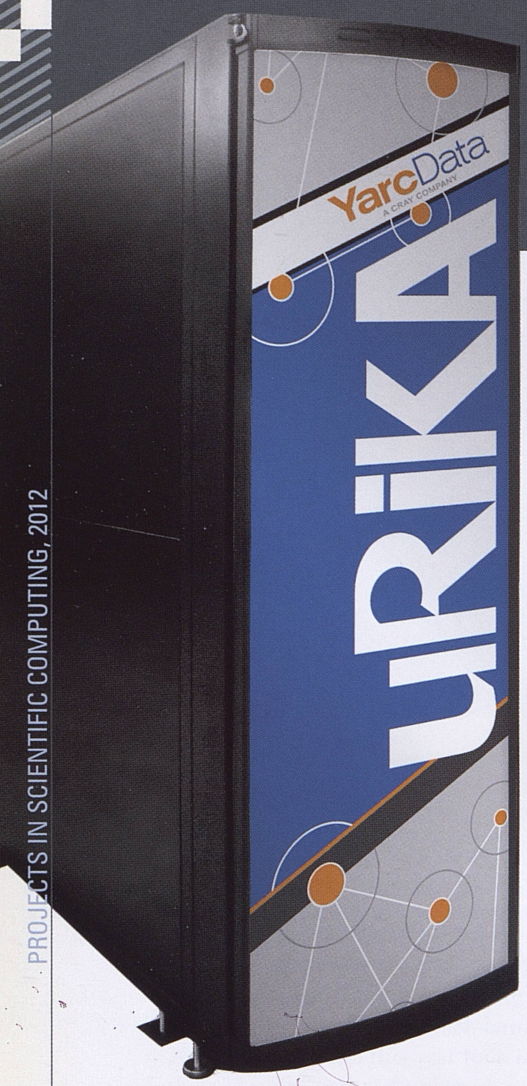
XSEDE

Extreme Science and Engineering Discovery Environment

**XSEDE Partners**
University of Illinois at Urbana-Champaign
Carnegie Mellon University & the University of Pittsburgh
University of Texas at Austin
University of Tennessee, Knoxville
University of Virginia
Shodor Education Foundation
Southeastern Universities Research Association
University of Chicago
University of California San Diego
Indiana University
Jülich Supercomputing Centre
Purdue University
Cornell University
Ohio State University
University of California, Berkeley
Rice University
The National Center for Atmospheric Research

More information: https://xsede.org

# Sherlock: Unlocking the Secrets of Big Data

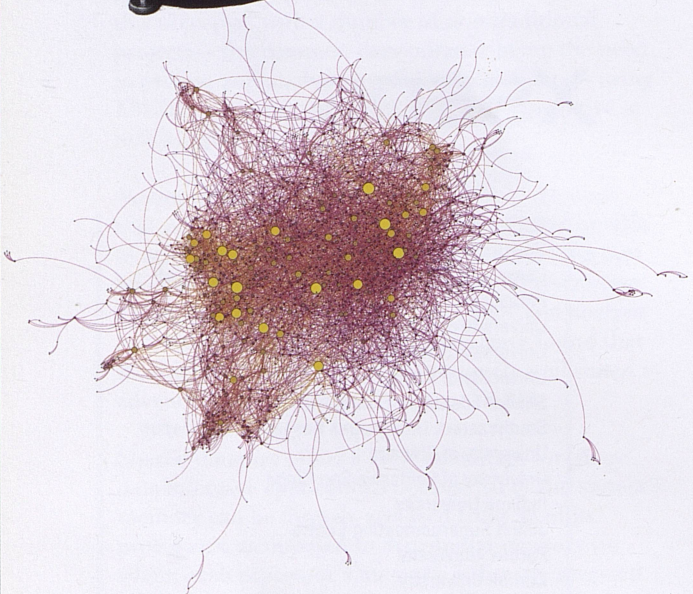



A YarcData uRiKA system representative of PSC's Sherlock

Computational analysis that discovers underlying patterns in "big data" can open many doors to understanding, such as how genes work, the dynamics of social networks, and the source of breaches in computer security. With this kind of analysis, based on a mathematical approach called "graph theory," interconnected webs of information can be represented as graphs, wherein nodes represent data elements and edges represent relationships among them.

Such graphs produced from real-world data can be huge, containing billions or trillions of edges. Even more challenging, these graphs typically can't be partitioned; their high connectivity prevents dividing them into subgraphs that can be practically mapped onto distributed-memory computers. "Graph analytics are notoriously difficult," says Nick Nystrom, PSC's director of strategic applications, "because following unpredictable paths from node-to-node is rate-limited by latencies to remote and local memory, which has drastically limited the graph problems that can be tackled."

To break the barrier blocking large-scale graph analytics, PSC this year introduced Sherlock, a unique supercomputer specialized for complex analytics on big data, which will be used for pilot projects by the national research community.



Protein-protein interactions in yeast, forming a relatively small graph of only 7,182 edges, illustrate the complexity of problems in graph analytics. (See Vladimir Batagelj & Andrej Mrvar (2006): Pajek datasets, http://vlado.fmf.uni-lj.si/pub/networks/data)

## Sherlock: The Details

Acquired through NSF's Strategic Technologies for Cyberinfrastructure program, Sherlock is a YarcData uRiKA ("Universal RDF Integration Knowledge Appliance") data appliance. It features massive multithreading, shared memory, and hardware optimizations to enable exceptionally efficient execution of graph algorithms. Sherlock contains 32 next-generation Cray XMT nodes. Aggregate shared memory is one terabyte, which can accommodate a graph of approximately 10 billion edges.

PSC customized Sherlock via additional Cray XT5 nodes having AMD Opteron processors to add valuable support for heterogeneous applications that use the XMT nodes as accelerators for graph-based algorithms. This heterogeneous capability will enable an even broader class of applications, for example in genomics, astrophysics, and other types of analysis of complex networks.

Sherlock runs an enhanced suite of familiar semantic web software for easy access to powerful analytic functionality, using the Resource Description Framework (RDF) as a very general and expressive data format. Sherlock also supports common programming languages such as C, C++, Java, Fortran, and scripting languages.

# Supercomputing in Pennsylvania

**PSC provides education, consulting, advanced network access and computational resources to scientists and engineers, teachers and students across the Commonwealth of Pennsylvania**

## 3ROX:Network for Education

The Three Rivers Optical Exchange (3ROX) (see pp. 12-13) provides research and education network service to seven Intermediate Units in western Pennsylvania that serve 116 school districts, more than 600 schools, 21,000 teachers and 300,000 students. 3ROX links these schools, teachers and students to a global community of people and ideas.

## Research & Training in Pennsylvania

Researchers in Pennsylvania are using PSC's new disk-based data storage, the Data Supercell (see p. 4), implemented with support from the Commonwealth of Pennsylvania's Redevelopment Assistance Capital Program. Pennsylvania organizations using the Data Supercell include the National Energy Technology Laboratory, Carnegie Mellon University, the Software Engineering Institute, the University of Pittsburgh Developmental Biology Program, and Drexel University's Design Arts Group.

Continuing a long-standing relationship with Lehigh University, PSC in March did a half-day workshop on parallel programming of multi-core computing systems. PSC scientist John Urbanic presented material on programmer-friendly standards (OpenMP and Open ACC) to 35 students as part of Lehigh's annual HPC Symposium.

In June, PSC scientific co-director Ralph Roskies and PSC director of networking Wendy Huntoon addressed information officers and other leaders of the 14 universities of the Pennsylvania State System of Higher Education. Their presentation highlighted how computational science and cyberinsfrastructure are changing research and outlined possible collaboration between PSC and PASSHE universities.

As part of a program sponsored by PAIUnet, Pennsylvania's statewide, high-speed educational network, PSC in March helped to develop and coordinate a data-modeling session for high-school students taking part in the state-wide Marcellus Shale Project. Also, through its BEST and CAST programs (see pp. 8-9), PSC provides continuing training and curriculum materials for western Pennsylvania high-school math and science teachers.

## Pennsylvania Research Innovation

Several projects in this booklet highlight research in Pennsylvania enabled through PSC:

- **Bright Lights, Big Cosmos:** Astrophysicists at Carnegie Mellon University are simulating the period in the evolution of the Universe when stars, galaxies and black holes first appeared (p. 40).
- **Modeling Aortic Aneurysms:** Drawing on data from Allegheny General Hospital, biomedical engineers are modeling aortic aneurysms so that it will be possible to better guide decisions on when surgery is required (p. 44).
- **When Small Worlds Collide:** A Lehigh University physicist is calculating spin properties of molecules that could help lead to quantum computing, much faster than today's supercomputers (p. 45).
- **Fighting Dengue Resurgence:** Researchers at PSC and the University of Pittsburgh are developing tools to help public-health decision makers intervene effectively to stop the world-wide spread of dengue fever (p. 47).

### Shared Memory Poker

Using PSC's Blacklight system, Carnegie Mellon computer science professor Tuomas Sandholm and his Ph.D. student Sam Ganzfried did well at Toronto in July — the Advancement of Artificial Intelligence (AI) annual Computer Poker Competition. In recent years, poker has emerged as an AI challenge similar to that served for many years by chess, but more demanding. "In poker," says Sandholm, "a player doesn't know which cards the other player holds or what cards will be dealt in the future. Such games of incomplete information are much harder to solve than complete-information games."

Sandholm's field, game theory, in which his work is internationally recognized, describes conflict in which the payoff is affected by actions and counter-actions of intelligent opponents. Like many games, poker can be formulated mathematically, but the formulations are unimaginably huge. Two-player no-limit Texas Hold'em poker has a "game tree" of about $10^{71}$ nodes, hence the usefulness of large amounts of memory. At Toronto, running with Blacklight, the Sandholm group's poker-playing agent finished second in the instant runoff scoring for two-player no-limit Texas Hold'em.

**Supercomputing Provided to Pennsylvania Organizations**

From July 2011 through June 2012, PSC provided more than 7.8 million processor hours to 917 individual Pennsylvania researchers from 40 institutions. The following Pennsylvania corporations, universities, colleges and K-12 institutions used PSC resources during this period:

*Albright College*
*Allegheny General Hospital*
*Allegheny-Singer Research Institute*
*Bryn Mawr College*
*Carnegie Mellon University*
*Cedar Crest College*
*Cheyney University of Pennsylvania*
*Community College of Allegheny County*
*Dickinson College*
*Drexel University*
*Duquesne University*
*Dynamix Technologies*
*Frazier School District*
*Grove City College*
*Haverford College*
*Indiana University of PA, all campuses*
*Kutztown University of Pennsylvania*
*Lehigh University*
*Life Technologies*
*Lock Haven University*
*Marconi Services*
*Oakland Catholic High School*
*Our Lady of Sacred Heart High School*
*PA CYBER Charter School*
*Pennsylvania State University, all campuses*
*Pittsburgh Public Schools*
*Pittsburgh Supercomputing Center*
*Shippensburg University of Pennsylvania*
*Slippery Rock University*
*Swarthmore College*
*Temple University*
*Thomas Jefferson University*
*University of Pennsylvania*
*University of Pittsburgh, all campuses*
*University of the Sciences in Philadelphia*
*Upper St. Clair High School*
*Ursinus College*
*Vitaerx*
*Wilkes University*
*Winchester-Thurston School*

# Energizing Science Learning

PSC programs in science education build bioinformatics expertise at minority-serving institutions and help to jumpstart the Pittsburgh region toward a cyber-savvy workforce

## Science Training for Faculty, Grad Students & Undergrads

MARC: "We've implemented a multi-disciplinary course in sequence-based bioinformatics at more than 10 universities," says PSC scientist Hugh Nicholas, who directs PSC's Minority Access to Research Careers (MARC) program. As of 2011, with NIH renewal of the program for five years, Nicholas and his colleague Alex Ropelewski are building a concentration or minor in bioinformatics at five partner minority-serving institutions (MSIs): North Carolina A&T; University of Puerto Rico, Mayaguez; Johnson C. Smith University; Tennessee State University; and Jackson State University.

Since 2001, the MARC program evolved from providing individual training in what was at first a newly emerging discipline, bioinformatics, to focus on the development of curricula and research programs. "The program has shifted direction," says Ricardo González, a professor at the University of Puerto Rico, School of Medicine, and co-principal investigator with Nicholas. "We've become good enough to establish bioinformatics programs or tracks at these universities and to provide a solid foundation for their faculty and students to carry out research in this field."

Many peer-reviewed papers have already resulted from the program, notes González, especially important in light of a 2011 study (*Science*, August 19, 2011) finding that black scientists were a third less likely than white counterparts to get a research project funded. "The PSC program has been addressing the uneven playing field that affects black researchers for 10 years."

Along with workshops at PSC, the MARC staff travels to partner MSIs to offer intensive on-site workshops. The program also provides a model bioinformatics curriculum, with course materials in related aspects of biology, computational science and mathematics, and offers teaching assistance for newly established courses. "At each campus with which we've partnered," says Nicholas, "we've trained people who are now capable of teaching a basic bioinformatics course."

Since 2003, the program has included a two-week workshop at PSC, the MARC Summer Institute, that trains graduate students – who can use bioinformatics tools for dissertation research – and faculty who plan to establish an introductory bioinformatics course at their home institution. The program also offers a 10-week internship at PSC, with nine participants this year. These internships build connections for young scientists with resources at two major research universities, Carnegie Mellon and the University of Pittsburgh, and have often led to published research.

*More information: http://marc.psc.edu*


## Training for High School Teachers in Science & Math

"Introducing 'cool' technology into the classroom engages students," says PSC's director of education, outreach and training, Cheryl Begandy, "and increases their willingness to stay with subjects they may otherwise find too complicated or just uninteresting." For Begandy and Pallavi Ishwad, education program director of PSC's National Resource for Biomedical Supercomputing (NRBSC), the goal is to help re-define high-school science instruction, so that it can better prepare future scientists, engineers and educators to participate in the cyber-savvy 21st-century marketplace.

The PSC EOT team (l to r): Robin Scibek, Debra Nigra, Cheryl Begandy (director), Vivian Benton and Pallavi Ishwad

BEST: Begun in 2007 by Ishwad, Better Educators of Science for Tomorrow (BEST) introduces high-school teachers to a bioinformatics curriculum adapted from PSC's MARC program for undergrad and graduate science students. Drafted and improved through classroom usage by an interdisciplinary group of STEM teachers, the BEST curriculum offers ready-to-use lesson plans for single-subject educators to extend their skills to the multidisciplinary outlook of bioinformatics, which draws on physics, chemistry, biology, computer science and math.

From June 15 to 21, Ishwad held a BEST workshop for teachers from the PA Cyber Charter School and from MARC partner MSI North Carolina A&T. She also extended BEST outreach efforts through high-school outreach units at other MARC partner MSIs, including Jackson State and Tennessee State.

Participants are learning tools of bioinformatics at the MARC 2012 Summer Institute.

"You have provided a tremendous amount of expertise and guidance in helping to shape our program," said Edwina Kinchington, of the Pittsburgh Science & Technology Academy, one of six southwest Pennsylvania high schools that have adopted BEST curricula as part of permanent elective course offerings. "The gift of this program to students is immeasurable," said biology teacher Rebecca Day of Frazier High School.

*More information on BEST: http://www.psc.edu/eot/k12/best.php*

CAST: In 2011, PSC received a $100,000 grant from the DSF Charitable Foundation that extends Computation and Science for Teachers (CAST), PSC's program – begun in 2008 – that has introduced many Southwest Pennsylvania STEM teachers to easy-to-use modeling and simulation tools for classroom learning.

The DSF grant funds a three-way effort among PSC and the Maryland Virtual High School Project, which helped to pioneer the use of computational thinking in high-school learning, and the Math & Science Collaborative of the Allegheny Intermediate Unit, which provides educational services to Allegheny County's 42 suburban school districts.

Educators from these organizations, with PSC providing overall direction and management, developed a Professional Development Program (PDP), an integrated set of modules to train teachers in western Pennsylvania in how to incorporate computational reasoning and the tools of modeling and simulation into math and science curricula.

The PDP's Depth track, the second of two tracks, was piloted during the CAST 2012 Summer Institute, from August 6-9. "CAST," says PSC's Begandy, "brings to the classroom the same problem-solving, technology-rich approaches currently used in scientific research and in business."

*More information on CAST: http://www.psc.edu/index.php/cast*

## Open Education Resources

Two PSC educational programs, CMIST and SAFE-Net, provide open education resources on the World Wide Web for educators, students and parents. SAFE-Net's website provides free materials to help parents, educators, students and individuals understand questions of cyber-security associated with wide usage of the Internet.

Through NRBSC, PSC also provides modules and vivid 3D video animations developed through its CMIST program (Computational Modules in Science Teaching). Three CMIST modules are available through the website: Molecular Transport in Cells; Big Numbers in Small Spaces: Simulating Atoms, Molecules and Brownian Motion; and Enzyme Structure and Function.

*SAFE-Net (free materials): http://safenet.3rox.net*

*CMIST (free modules): http://nrbsc.org/cmist*

# The National Resource for Biomedical Supercomputing

## National Leadership in High-Performance Computing for Biomedical Research

▲ **The NRBSC team:** (l to r) Markus Dittrich (director), Greg Hood, Hugh Nicholas, Pat Sudac, Art Wetzel, Alex Ropelewski, Pallavi Ishwad. (Not in photo: Jacob Czech)

Established in 1987, PSC's National Resource for Biomedical Supercomputing (NRBSC) was the first external biomedical supercomputing program funded by the National Institutes of Health (NIH) and has continued uninterrupted since then. Along with core research at the interface of supercomputing and the life sciences, NRBSC scientists develop collaborations with biomedical researchers around the country, fostering exchange among experts in computational science and biomedicine and providing computational resources, outreach and training.

In September 2012, NRBSC gained renewed NIH support as a participant in a newly funded Biomedical Technology Research Center (BTRC) on High-Performance Computing for Multiscale-Modeling of Biological Systems. This $8 million grant, from NIH's National Institute of General Medical Sciences, establishes a collaboration between the University of Pittsburgh, Carnegie Mellon University and PSC. The principal investigator of the BTRC, Ivet Bahar, chairs the Department of Computational & Systems Biology at the University of Pittsburgh School of Medicine. Markus Dittrich of NRBSC leads the PSC component, and Robert F. Murphy, director of the Lane Center for Computational Biology, leads Carnegie Mellon's participation.

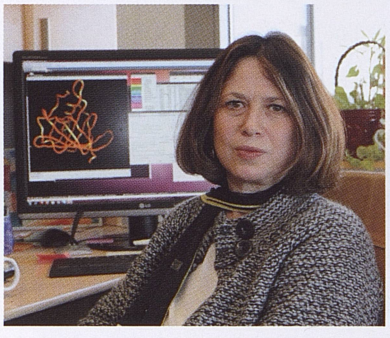
"This collaboration opens many opportunities as NRBSC goes forward," says Dittrich. "Through the new BTRC we continue our work in cellular modeling, structural biology, and large-scale volumetric image analysis, and we gain through synergy with the outstanding computational biology programs at the University of Pittsburgh and Carnegie Mellon."

### Anton Program Extended

A supplementary award to the newly established BTRC also provides $1.1 million to extend the Anton program (see pp. 26-31) for another two years. In partnership with D. E. Shaw Research (DESRES), this program makes an innovative computing system available to U.S. biomedical scientists. Having served 91 research projects by more than 70 individual research groups in two years, the Anton program commenced a new round of allocations in November 2012.

"We are thrilled about the impact that Anton has had over the last two years," says Markus Dittrich of NRBSC, "and we are excited to be able to offer continued access to this great resource for the biomedical community."

More info: http://www.nrbsc.org



**Ivet Bahar**, Professor and John K. Vries Chair, Department of Computational & Systems Biology, School of Medicine, University of Pittsburgh, leads the newly established Biomedical Technology Research Center on High-Performance Computing for Multiscale-Modeling of Biological Systems.

"I'm very optimistic about the collaborative possibilities of this joint effort," says Bahar. "Our vision is to begin to fill the gaps among modeling efforts at disparate scales of structural biology, cellular microphysiology and large-scale image analysis. We will build the framework to unify these efforts and gain insights that will help to alleviate neurobiological disorders. The experience and expertise of PSC are an essential element of this."

**Robert F. Murphy**, Ray and Stephanie Lane Professor of Computational Biology and Professor of Biological Sciences, Biomedical Engineering, and Machine Learning, Carnegie Mellon University. He was a co-principal investigator on the original PSC biomedical supercomputing grant.

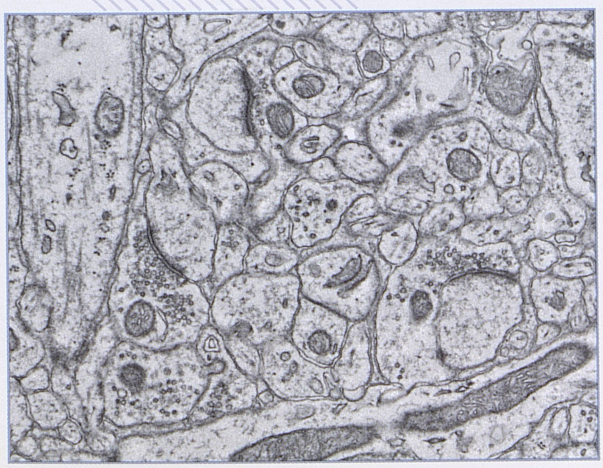
"We have imagined this new center as a Pittsburgh center, joining the two universities, the University of Pittsburgh and Carnegie Mellon, with PSC strengths in training and in spatially realistic cellular modeling, structural biology, and large-scale volumetric image analysis. We now have an opportunity to combine that work with work in the Lane Center on image-derived modeling of cellular organization and machine learning for structural biology and to go beyond what we've done before."

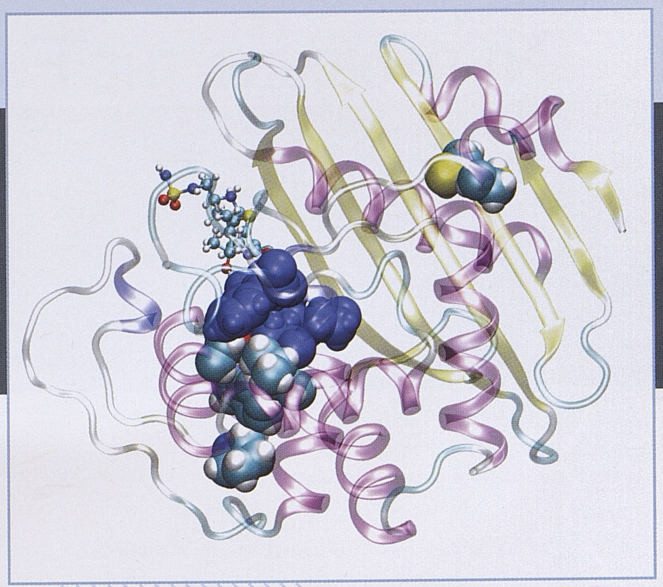
## Research

**NRBSC research focuses on three areas of biomedicine that span many scales of space and time:**

**Spatially realistic cell modeling** centers on stochastic computer simulations of movements and reactions of molecules within and between cells, to better understand physiological function and disease. MCell and CellBlender software is developed at the NRBSC and used to model and visualize events such as (shown in this image) neurotransmission between nerve and muscle cells.

**Biomedical image processing** using NRBSC developed software enables accurate three-dimensional reconstruction of brain circuits from massive serial-section electron microscopy image sets. The example shown here spans a cortical region of mouse brain and is built from 3000 individual camera images at four nanometer resolution. After registration at the NRBSC, the large sectional reconstruction occupies 12 gigabytes. Thousands of these sections are then combined and aligned into 3D volumes for visualization and analysis of neural pathways.





**NRBSC structural biology** focuses on computational tools to determine the structure of proteins from their amino-acid sequence and quantum-mechanical simulation methods for biomolecules such as enzymes. This image shows structure for the oxacillinase class D beta-lactamase enzyme, with the "active site" in blue. Beta-lactamase is a bacterial enzyme that is a major mechanism of antibiotic resistance. PSC-developed software enables researchers to simulate enzyme reactions and gain new insight into enzyme function, which facilitates design of new therapeutic drugs.

### PSC's NRBSC Workshops (2011-2012)

Computer Simulation of Biomolecular Dynamics and Reactions

Computational Methods for Spatially Realistic Microphysiological Simulations

Summer Institute in Bioinformatics
(for minority-serving institutions)

Anton Training Workshop

Bioinformatics Internship Program

*NRBSC and PSC have also developed educational programs, CMIST and BEST (see pp. 8-9), that have provided training to high-school and undergrad students and educators in the Pittsburgh region and nationally.*

**NRBSC Biomedical Collaborations**

Albert Einstein College of Medicine
Carnegie Mellon University
Duke University
Harvard University
Howard University
Howard Hughes Medical Institute, Janelia Farm Research Campus
Allen Institute for Brain Science
Grand Valley State University
The Salk Institute
University of California at Davis
University of California at San Diego
University of Pittsburgh
University of Pittsburgh School of Medicine
University of Puerto Rico, Medical Sciences Campus
University of Michigan

# Networking the Future

## One of the leading resources in the world for network know-how

PSC's Advanced Networking group is one of the leading resources in the world for knowledge about networking. Through 3ROX (Three Rivers Optical Exchange), a high-speed network hub, they operate and manage network infrastructure that connects many universities and schools in Pennsylvania and West Virginia to research and education networks, such as Internet2 and National LambdaRail (NLR), that link to universities, corporations and research agencies nationally. Their research has created valuable tools for improving network performance. In a current project, Web10G, PSC network staff are helping to develop software to enable non-expert users to more fully exploit the bandwidth of advanced networks.

*More information: http://www.psc.edu/networking/*

### Improved Connectivity in Pennsylvania and West Virginia

In January, 3ROX and Drexel University opened a high-performance network link, Philadelphia to Pittsburgh, via high-performance, fiber-optic network. This link, using the FrameNet service of NLR, provides bandwidth of 10 Gigabit Ethernet (10 billion bits per second), about 100 times faster than current high-end download rates for most residential Internet service.

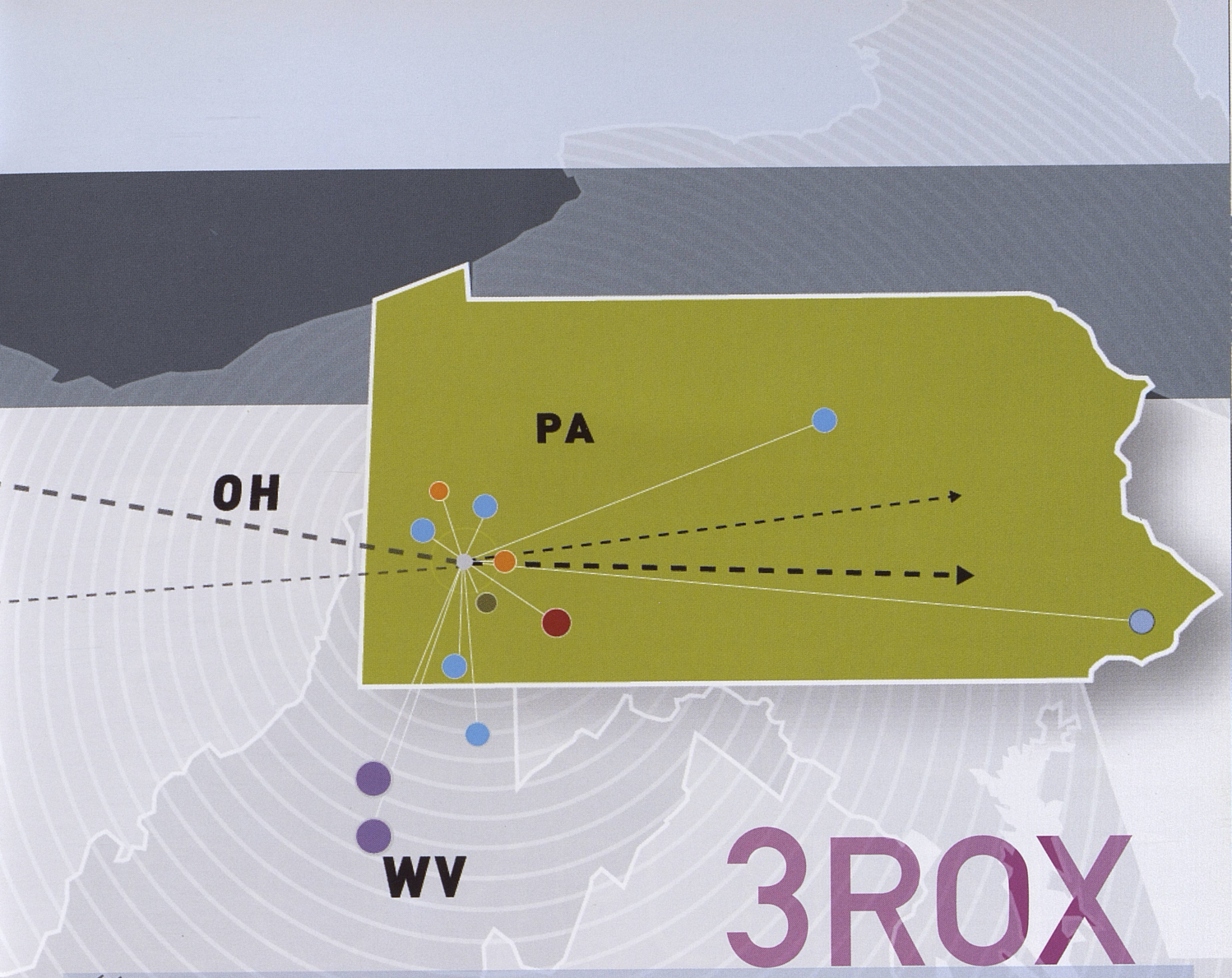
The new link augmented connectivity through the 3ROX/Drexel Internet2 hub. This partnership, formed in early 2011, provides network services to universities, research sites and K-12 schools in western Pennsylvania and West Virginia along with Drexel and its affiliated research sites and 14 Pennsylvania State System of Higher Education universities. "This link," said Huntoon, "is the framework for a consolidated high-performance network infrastructure across Pennsylvania."

This year 3ROX also improved network service in West Virginia, upgrading the link between PSC and West Virginia University (WVU) and adding WVNET (West Virginia Network) as a member. The new link to WVU increases bandwidth 64-fold. "This is a big step forward for research and education connectivity to WVU," said Huntoon.

By joining as a participant in 3ROX, WVNET upgrades connectivity from West Virginia K-20 schools to research and education networks such as Internet2 to two 10 gigabit-per-second connections (one from Morgantown and another from Huntington). More important than the improved bandwidth per se, says Dan O'Hanlon, director of WVNET, is the collaboration with 3ROX. "It's a big gain for West Virginia," said O'Hanlon, "that we're now able to collaborate with people who are involved in supercomputing and have world-class experience in running a research and education network."

The catalyst for the upgrades, says Huntoon, was the 10 Gbps connection that 3ROX provided last year for the National Oceanic and Atmospheric Administration Environmental Security Computing Center in Fairmont, West Virginia. "The NOAA grant was stimulus funding," says Huntoon, "and because we had infrastructure in place, we've been able to provide these expanded services very competitively from a cost perspective. These upgrades transform the West Virginia Internet landscape."

▲ **Wendy Huntoon**, PSC director of networking, has served in several national leadership roles in research and education networks and is currently chief architect in the office of the chief technology officer for Internet2, an advanced networking consortium led by the research and education community.




# 3ROX

**3ROX members**

- **UNIVERSITIES**
  Carnegie Mellon University, Pennsylvania State University, Robert Morris University, University of Pittsburgh, Waynesburg University, West Virginia University.
- **K-12 INSTITUTIONS**
  Allegheny Intermediate Unit (AIU3), Arin Intermediate Unit (IU28), Beaver Valley Intermediate Unit (IU27), Intermediate Unit One, Northwest Tri-County Intermediate Unit (IU5), Riverview Intermediate Unit (IU6), City of Pittsburgh School District (IU2), Seneca Highlands (IU9).
- **GOVERNMENT LABORATORIES AND FACILITIES**
  The National Energy Technology Laboratory; NOAA Environmental Security Computing Center.
- **BUSINESS**
  Westinghouse Electric Co.
- **OTHER**
  Computer Emergency Response Team, WVNET.
- **RESEARCH NETWORK PARTNER:** *Drexel University*

**Network Connections**

- **NATIONAL RESEARCH NETWORKS**
  Internet2 — 5 Gbps, ESnet — 1 Gbps, National LambdaRail PacketNet — 10 Gbps, XSEDE — 10 Gbps.
- **NATIONAL COMMODITY INTERNET NETWORKS**
  Level 3 — 10Gbps; Cogent — 1 Gbps.
- **PITTSBURGH LOCAL EXCHANGE NETWORKS**
  Comcast, MetNet, TeraSwitch & Cavalier.
- **OTHER NETWORK CONNECTIONS**
  Southern Crossroads (SOX) — 1Gbps, TransitRail-CPS — 4 Gbps, OARnet — 10 Gbps, FrameNet — 10 Gbps, WVNET — 10 Gbps.

Note: Gbps: a billion (Giga) bits per second.

# THE SUPER COMPUTING SCIENCE CONSORTIUM

## Pennsylvania-West Virginia partners in development of clean power technologies.

▲ (SC)² co-chairs Lynn Layman, PSC (right) & Bob Romanowsky, NETL.

Formed in 1999 and supported by the U.S. Department of Energy, the Super Computing Science Consortium is a regional partnership of research and educational institutions in Pennsylvania and West Virginia. (SC)² provides intellectual leadership and advanced computing and communications resources to solve problems in energy and the environment and to stimulate regional high-technology development and education.

Since the spring of 2000, a high-speed network – the first fiber-optic service to Morgantown, West Virginia – has linked the National Energy Technology Laboratory (NETL) campuses in Morgantown and Pittsburgh with PSC, facilitating NETL collaborations.

**(SC)² PARTNERS**
National Energy Technology Laboratory
Pittsburgh Supercomputing Center
Carnegie Mellon University
University of Pittsburgh
Waynesburg University
West Virginia University

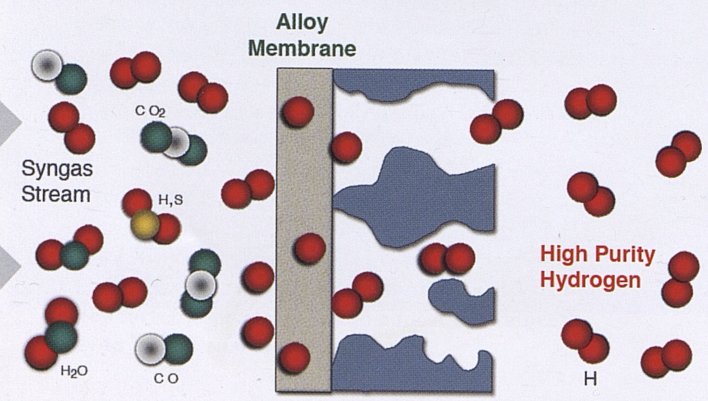
**More information**: http://www.sc-2.psc.edu

### PSC & (SC)²: Research for Clean Energy

Since the 1999 founding of (SC)², 55 (SC)² research groups, including 114 researchers, have used PSC systems for a range of clean-energy related projects, including designs for advanced power turbines, fluidized-bed combustion, and a reactor to produce power from gasified coal. This work has used more than 6.8 million hours of computing time, over 285,000 hours during the past year.

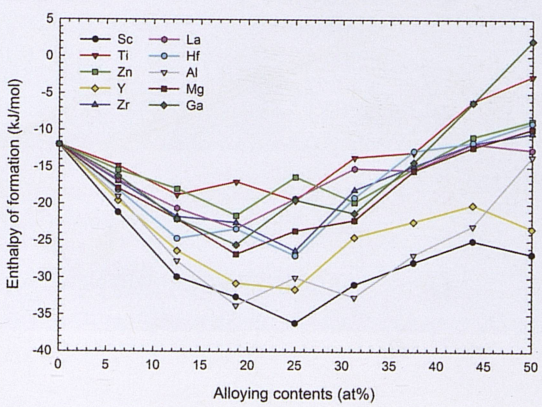
### Pure Hydrogen from Coal Gas

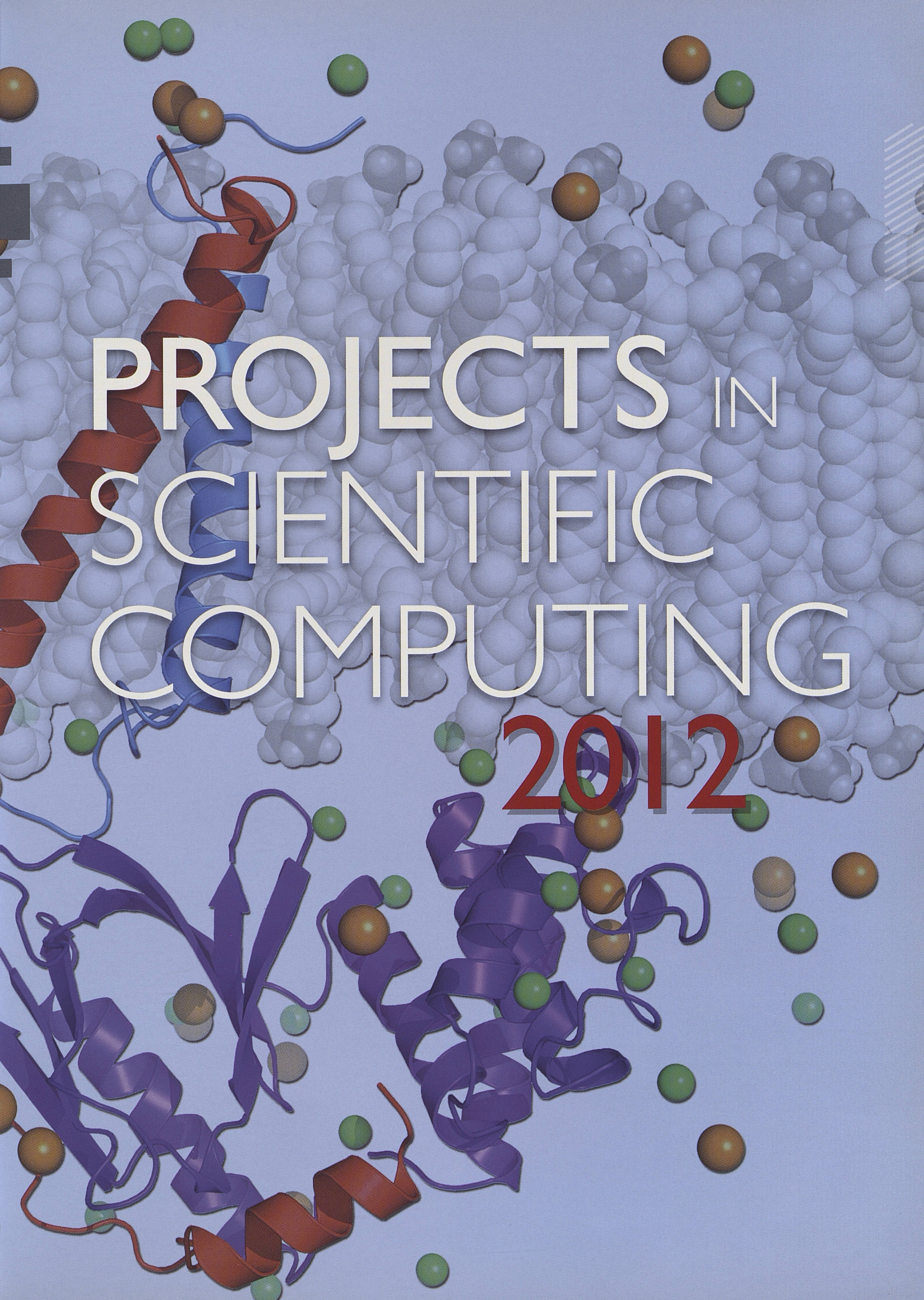
Membranes aren't only ear drums and cell walls. NETL researchers are developing advanced membrane technologies that can, among other things, separate pure hydrogen from the "syn gas" produced in reactors that convert coal into gases. The high-purity hydrogen that results is a valuable clean-energy commodity – for hydrogen fuel cells and many other industrial uses. The overall goal is to improve the efficiency and robustness of these membranes, conventionally metallic alloys of copper and palladium, so that they can be efficiently integrated into coal-conversion processes.




Materials scientist Michael Gao of NETL (and URS Corporation) has used PSC'S Blacklight to accelerate design and development of new, improved membrane alloys of copper-palladium with a third element (ternary alloys) through quantum-mechanical simulations. One project (illustrated by the graph) involves calculating the "enthalpy" – a thermodynamic property – of proposed compounds with different alloying elements.

"We look at the periodic table," says Gao, "and take atoms to put into virtual alloys. These computations allow us to test hundreds of possibilities and select the most viable candidates to test in the laboratory. We accelerate materials development and shorten the time to improved hydrogen separation performance at reduced cost." Gao collaborates in these studies with Bryan Morreale's group in NETL-Pittsburgh, Andrew Gellman's group at Carnegie Mellon University, and Ömer Doğan's group in NETL-Albany.





# PROJECTS IN SCIENTIFIC COMPUTING

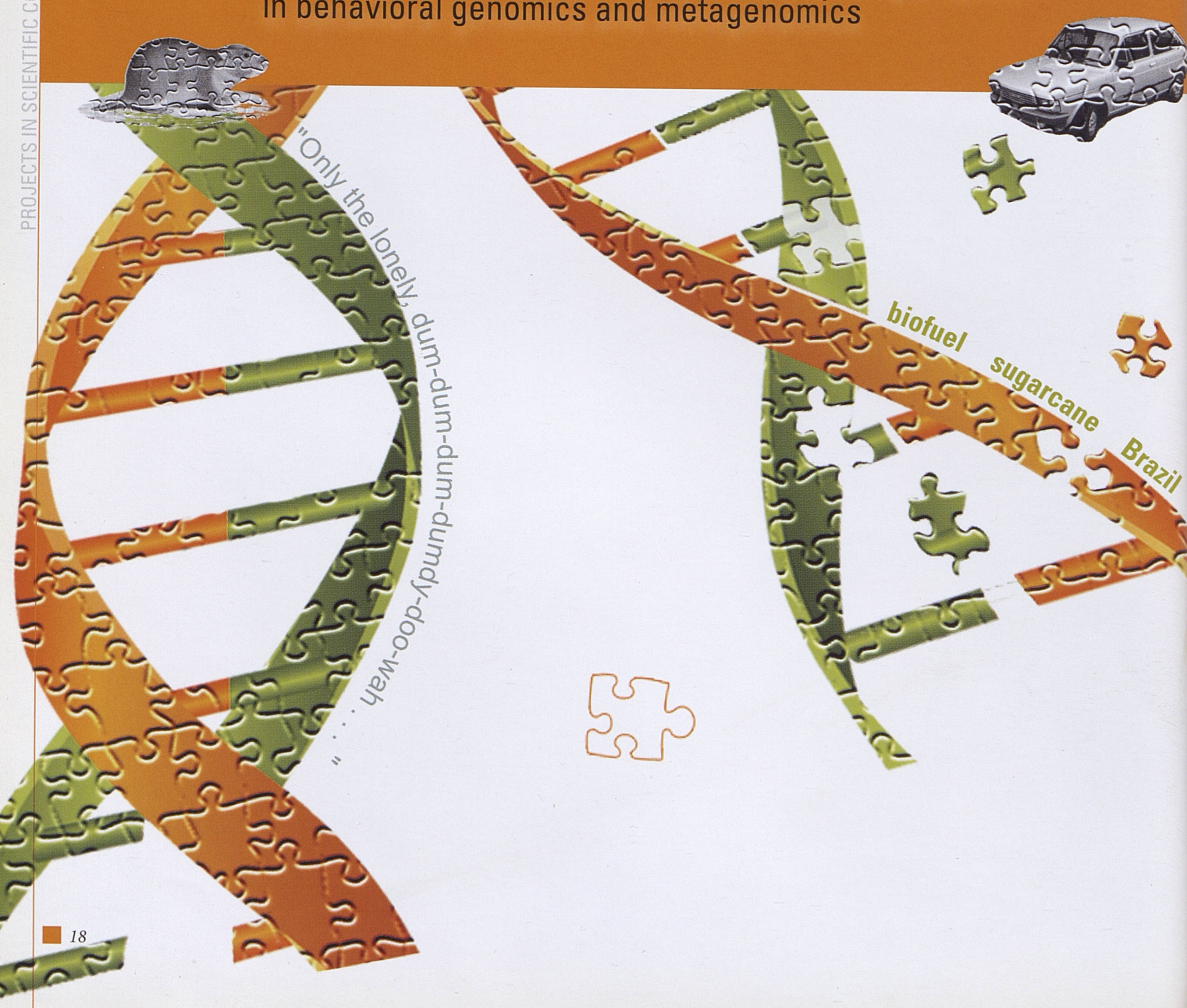
## 2012

# PROJECTS 2012
# CONTENTS

## SHARED MEMORY GENE

# ASSEMBLY

Because of its shared-memory design, PSC's Blacklight has become an invaluable tool in assembly and analysis of next-generation sequencing data, including projects in behavioral genomics and metagenomics



**Phil Blood**, PSC, XSEDE Extended Collaborative Support Services

In genomics, the next generation is now. This relatively new branch of the life sciences has in the last few years, due to new technologies, exploded with possibility and data. "Next generation" sequencing tools have taken genomics well beyond the Human Genome Project to studies of nearly every kind of organism, from ants and bumblebees to Patagonian tuco-tucos (more on that below) among many others, by deciphering the order of nucleotide bases – A, G, C and T (adenine, guanine, cytosine and thymine) – at unprecedented speed.

The essential difference is long versus short reads. Previous sequencers did reads of about 300 to 500 and sometimes up to 1000 bases. The new technologies gain their advantage by doing much shorter reads, 50 to 150 bases – at greatly reduced cost per base – and can generate in a week as much sequence data as would require a year for the traditional sequencers. Consequently, genomics has shifted into data-intensive overdrive, with many opportunities to do important research. While it's an unprecedented blessing for the life sciences, it's also an unprecedented challenge for data processing and analysis.

Once a sequencing instrument has produced millions or, as the case may be, billions of reads from an organism's DNA (or RNA), researchers face the task of assembling them. To add to the degree of difficulty, short reads amplify the computational challenge – many more pieces of data must be fit together based on shorter overlaps. How do you assemble all those sequence fragments into complete and accurate genomic strands? Imagine a jigsaw picture puzzle with 100 big pieces versus the same picture with 2,000 little pieces. It's a potentially mind-boggling problem handled by very sophisticated computational algorithms, requiring many runs, careful checking and, some say, as much art as science.

Since Blacklight came online in October 2010, with two partitions of 16 terabytes of shared memory, the largest shared-memory system in the world, it's become a powerful tool in meeting this challenge. An article in *GenomeWeb* (February 1, 2012) highlighted Blacklight's advantages – very large shared memory making it possible to contain entire base-pair datasets in random-access memory (RAM) – dramatically improving workflow and throughput time, as compared to non-shared memory clusters, for genomics assembly and analysis.

Beyond that, observes PSC scientist and XSEDE consultant Phil Blood, many large genomics assemblies simply couldn't be done without large shared-memory, such as Blacklight provides. To enhance Blacklight's genomics capabilities for XSEDE researchers, Blood has made nearly all genomics software tools available for easy use. "Blacklight has an extensive collection of pre-compiled modules for the analyses of next-generation sequence data," says Matthew MacManes, of the University of California, Berkeley. MacManes attempted genomics analysis with a number of high-end computing systems before coming to Blacklight. He used nearly 20 different programs in his assembly and analysis at PSC: "Having these programs installed and maintained by PSC staff is extremely helpful."

MacManes' project with Blacklight involved assembly and analysis of RNA from tuco-tucos, a burrowing rodent from Patagonia – of particular interest in that some tucos live in social groups while others of the same species are anti-social and live alone. MacManes identified a number of genes that are expressed or not depending on whether the tucos live alone or in a colony. "For my research," says MacManes, "the use of Blacklight has been absolutely revolutionary, allowing me to complete analyses that link specific patterns of gene expression with mammalian social behavior."

In science, new tools often make it possible to look at new questions, and the availability of next-generation sequencing has led to studies in "metagenomics" – analysis of genes from many organisms that co-exist in the same place – unimaginable a few years ago. "In metagenomics, unlike the traditional approach of analyzing the genome of one organism, you can pick an environment and take a sample," says Blood. "It could be Old Faithful, or what's inside human intestines, wherever you might find interesting microbial communities."

In a metagenomics study with Blacklight, Blood helped researchers from Oklahoma State assemble sequencing data from soil that came from a sugar-cane plantation in Brazil. The goal is to find enzymes that can efficiently break down non-feed plants, such

as switchgrass, wheat straw and others that have the potential to yield biofuel more efficiently than feed-stock plants like corn. Thanks to Blacklight, the Oklahoma State team – Mostafa Elshahed, Rolf Prade and Brian Couger, with Couger handling the computation – completed the largest metagenomics assembly to date.

"It wouldn't have been possible for us to do this on any other system," says Couger. Their work, still in analysis, has identified thousands of candidate enzymes, all previously unknown, that offer promise to cost-effectively degrade non-feed-stock crops to biofuel.

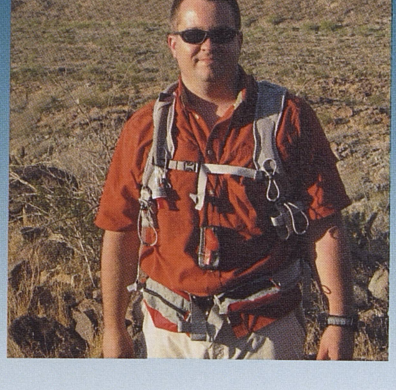
## Behavioral Genomics: Alone Time for Tucos

"Does behavior evolve through gene expression changes in the brain in response to environment?" The question, posed by a leading genomics scientist, Gene Robinson, caught MacManes' attention in 2009, when he was challenged by his laboratory group leader, Eileen Lacey, to come up with a dream research project. The answer to Robinson's question, says MacManes, is "yes, but . . . which genes, and how do we find them?"

Evolutionary biology explains that species maintain group behavior when the survival benefit is greater than cost – birds flock, for example, wolves run in packs, and tigers are territorial and mainly solitary. With these and many other examples, how do differences in social behavior show up in genes? Studies have identified a few genes that appear to be involved across a number of unrelated species, mostly insects, but there's little consensus, says MacManes, about the genetic underpinnings of social versus solitary behavior.

These thoughts led MacManes to what he calls his "craziest idea ever" – which now, only three years later, because of rapidly evolving genomics technologies, seems not so crazy at all. What if, he thought, he could look at the genomics in a case where both ends of the solitary versus social spectrum are represented in the same species?

Conveniently, such a species, the colonial tuco-tuco (*Ctenomys sociabilis*) was available, a population of them housed and studied at Berkeley's Museum of Vertebrate Zoology, where MacManes is an NIH-sponsored post-doctoral fellow. The colonial tuco-tuco – so-called for a clicking sound it makes – is a subterranean, burrowing rodent from Patagonia, related to the common guinea pig, unusual in the intra-species variation in behavior it exhibits. "Some of the females," says MacManes, "live in colonies with larger family groups, while others – at about one year of age – disperse from their birth burrow and live alone. Most social animals are obligately social – there aren't usually solitary animals to be found, and this variation makes tucos interesting and unique."



Matthew MacManes ▶

Working with two control populations of five tucos each, housed in social and solitary conditions, MacManes used messenger RNA from the hippocampus – a brain region implicated by prior research in social behavior. The extracted tissue was sequenced (in an Illumina sequencer), yielding 56 billion base pairs of raw data – 560 million 100 base-pair reads, "a ridiculous amount of reads," says MacManes.

***"Currently there is simply no better resource for this type of work."***

To grapple with assembly and analysis of this huge amount of sequence data, beginning with the task of building complete "transcriptomes" – full strands of RNA – MacManes first turned to large distributed-memory machines at several sites, but eventually came to PSC's Blacklight. Using 80 cores of Blacklight, 640 gigabytes of RAM, he completed the assembly in 14 days of computing, with subsequent analysis extending for months.

The work identified a number of genes that are differentially expressed depending on tuco social behavior, and MacManes and Lacey have a manuscript in preparation reporting their findings. "Blacklight is a key resource for my analyses of next-generation sequence data," says MacManes. "Without it, I would simply have been unable to complete the requisite analyses. I feel so strongly about Blacklight that I have referred colleagues and collaborators. Currently there is simply no better resource for this type of work."

▲ The colonial tuco-tuco (*Ctenomys sociabilis*) in its native environment (volcanic ash). A burrowing rodent, tucos are of interest in behavioral genomics because they vary, in the same species, between social and solitary living conditions.

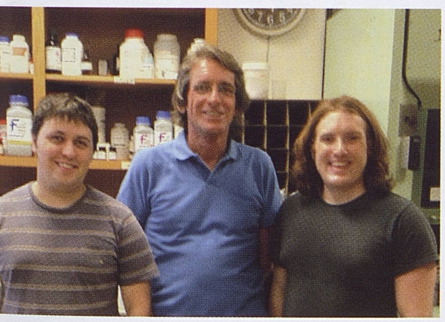
▲ The 1979 Brazilian Fiat 147 was the first modern automobile capable of running only on ethanol, which Brazil produces from sugarcane, shown ready for harvest at a plantation in São Paulo State.

## Metagenomics: Looking for Biofuel Enzymes

"In Brazil," says Rolf Prade, a professor of microbiology at Oklahoma State University and a Brazilian, "biofuel is standard gasoline for cars." The world's second largest producer of biofuel, Brazil gets its ethanol from sugarcane, uniquely available there due to enormous amounts of arable land and suitable climate.

Like corn in the United States, however, sugarcane has major disadvantages as a biofuel source. The amount of energy available per amount of input crop is much greater in non-feed stock plants, with denser fibrous structure (lignocellulosic plants), such as switchgrass and wheat straw, but these non-food plants are expensive and difficult to degrade into biofuel.

Considerable research worldwide is focused on finding inexpensive means to overcome this obstacle. A team at Oklahoma State led by Mostafa Elshahed, along with Prade and Brian Couger, has taken an innovative metagenomics approach. To find enzymes that can do the heavy-duty biodegrading, they started with a soil sample, gathered by Prade, from a Brazilian sugar-cane field.



▲ **Brian Couger** (right), **Rolf Prade** (center) and **Tyler Weirick**, Oklahoma State University

"These fields have been growing for 50 years," says Prade. "They cut the stems off the plants and throw everything else back in the soil and let it recycle – so this is very efficient recycling soil." To further accumulate microorganisms involved in biomass degradation, the researchers enriched the sample in a bioreactor with oxygen and more sugarcane biomass, and cultured the soil for eight weeks.

To isolate DNA, the researchers then sequenced the soil (in an Illumina next-generation sequencer). This yielded 1.5 billion pair-data reads of 100 base pairs each, approximately 300 gigabases in total. With Couger handling the computing, the researchers turned to Blacklight for the assembly, using software called Velvet – made available on Blacklight as a pre-compiled module by XSEDE consultant Blood of PSC.

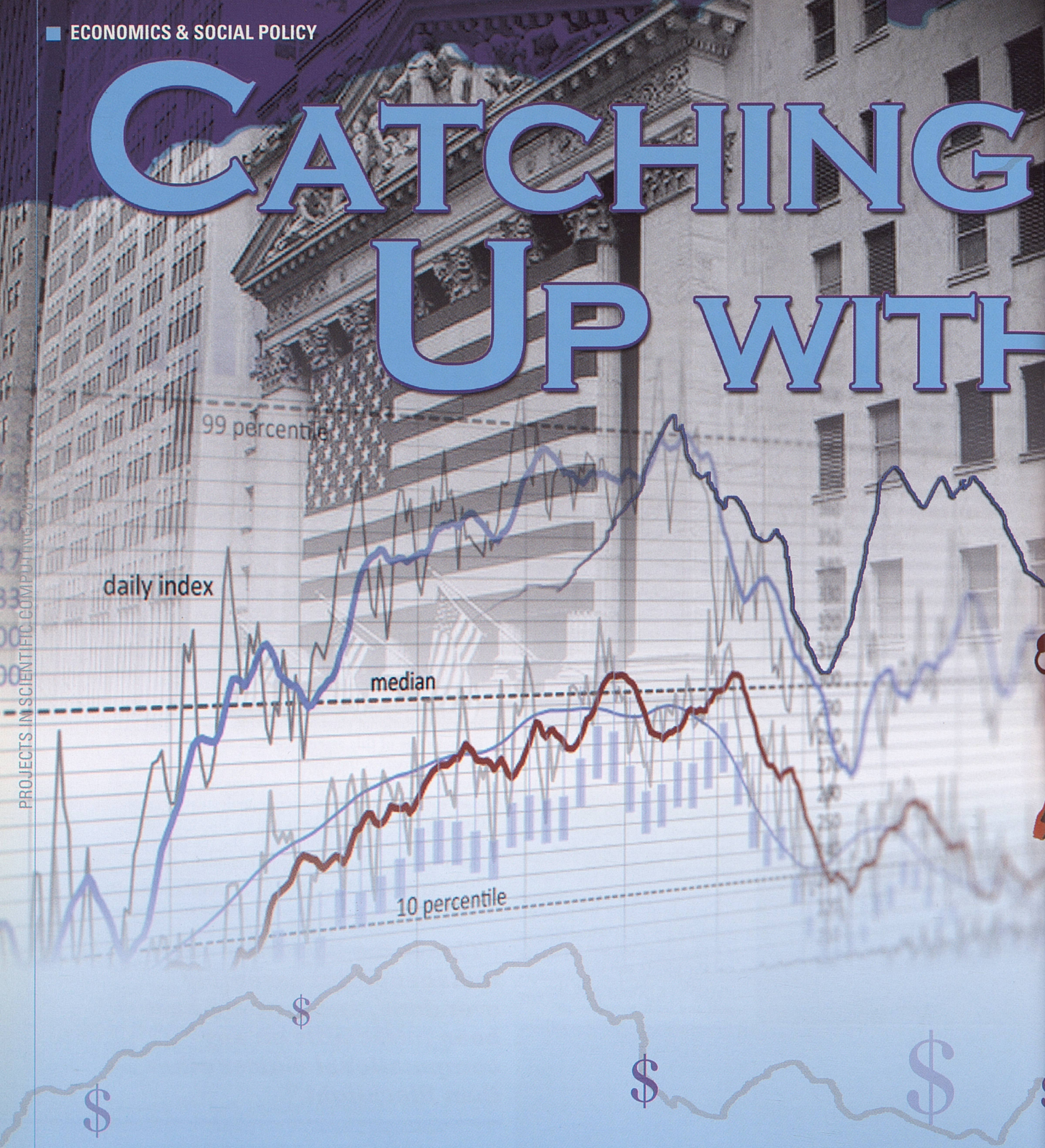
The entire metagenomics dataset occupied 3.5 terabytes of Blacklight memory. "This is the largest metagenomics assembly ever done," says Couger, "and it would have been intractable on any computational cluster other than Blacklight." To make use of this assembled data as a means to identify enzymes, the researchers also did a series of protein separations from the soil samples and applied mass spectrometry. From precise molecular weights, they inferred amino-acid sequences. Matching these sequences with the assembly data, the researchers identified more than 8,000 gene candidates related to glycoside hydrolase, a category of enzymes that can degrade plant cell walls.

***"This is the largest metagenomics assembly ever done, and it would have been intractable on any computational cluster other than Blacklight."***

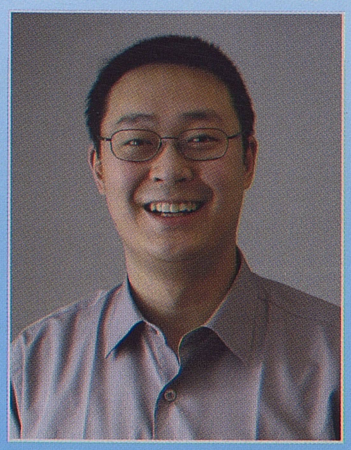
"This is like a protein discovery platform," says Prade. "Making biofuels from lignocellulosics doesn't work because we don't know how to decompose the biomass. It's because we don't have all the proteins, and we're working to find those." ▮

**More info:** www.psc.edu/science/2012/assembly/

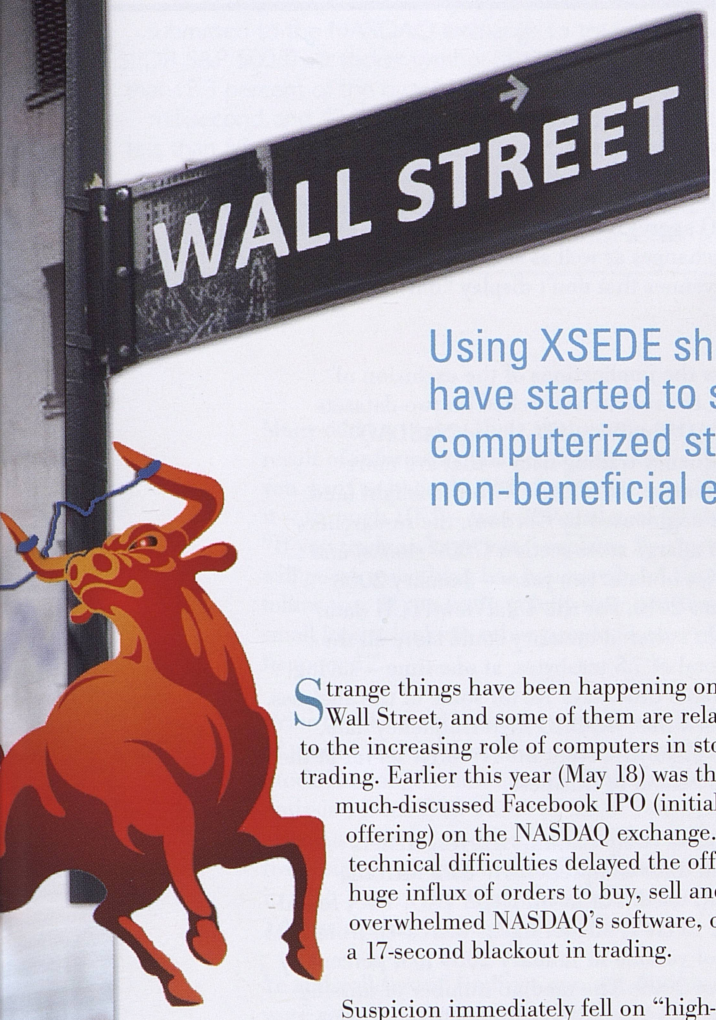
# Catching Up With





**Mao Ye**, University of Illinois, Urbana-Champaign. Ye credits XSEDE resources and consulting assistance. "Without XSEDE and shared memory, we wouldn't be able to effectively study these large amounts of data produced by high-frequency trading."




## Using XSEDE shared-memory resources, researchers have started to show how the rapid speed of computerized stock trading may have little understood, non-beneficial effects on the market

Strange things have been happening on Wall Street, and some of them are related to the increasing role of computers in stock trading. Earlier this year (May 18) was the much-discussed Facebook IPO (initial public offering) on the NASDAQ exchange. After technical difficulties delayed the offering, a huge influx of orders to buy, sell and cancel overwhelmed NASDAQ's software, causing a 17-second blackout in trading.

Suspicion immediately fell on "high-frequency trading" (HFT) – a catch-all term for the practice of using high-powered computers to execute trades at very fast speeds, thousands or millions per second. Since the U.S. Securities and Exchange Commission (SEC) authorized electronic trades in 1998, trading firms have developed the speed and sophistication of HFT, and over the last few years, it has come to dominate the market.

With HFT, profits accrue in fractions of a penny. A stock might, for instance, momentarily be priced slightly lower in New York than London, and with an algorithm in charge, an HFT trader can almost instantaneously buy and sell for risk-free profit. With HFT, traders typically move in and out of positions quickly and liquidate their entire portfolios daily. They compete on the basis of speed.

In June, *The Wall Street Journal* reported that trading had entered the nanosecond age. A London firm called Fixnetix announced a microchip that "prepares a trade in 740 billionths of a second," noted the *WSJ*, and investment banks and trading firms are spending millions to shave infinitesimal slivers of time off their "latency" to get to picoseconds, trading in trillionths of a second.

Does faster equal better? HFT has happened so quickly that regulators and academics are barely beginning to delve into the complex implications. In theory, increased trade volume and improved liquidity – the ease of buying and selling – makes markets more accurate and efficient. But HFT is a different beast from traditional investing, which places a premium on fundamental analysis, information and knowledge about businesses in which you invest.

Many questions arise about fairness and things that can go wrong (such as computer glitches) to the detriment of the market. One of the first problems researchers face, however, is that with HFT the amount of data has exploded almost beyond the means to study it – a problem highlighted by the "flash crash" of May 6, 2010. The Dow Jones Industrial Average dropped nearly 1,000 points, 9 percent of its value, in about 20 minutes, the biggest one-day drop in its history. Analysis eventually revealed HFT-related glitches as the culprit, but it took the SEC five months to analyze the data and arrive at answers.

"Fifteen years ago, trade was done by humans," says Mao Ye, assistant professor of finance at the University of Illinois, Urbana-Champaign (UIUC), "and you didn't need supercomputing to understand and regulate the markets. Now the players in the trading game are superfast computers. To study them you need the same power. The size of trading data has increased exponentially, and the raw data of a day can be as large as ten gigabytes."

To directly address the data problem and a number of other questions related to HFT, Ye and colleagues at UIUC and Cornell turned to XSEDE, specifically the shared-memory resources of Blacklight at the Pittsburgh Supercomputing Center (PSC) and Gordon at the San Diego Supercomputer Center (SDSC). Anirban Jana of PSC and XSEDE's Extended Collaborative Support Services worked with Ye to use these systems effectively.

In a study they reported in July 2011, the researchers – Ye, Chen Yao of UIUC and Maureen O'Hara of Cornell – processed prodigious quantities of NASDAQ historical market data, two years of trading – to look at how a lack of transparency in odd-lot trades (trades of fewer than 100 shares) may skew perceptions of the market. Their paper, "What's Not There: The Odd-Lot Bias in TAQ Data," was published in *The Journal of Finance*, the top journal in this field. The study has received wide attention, and in September, as a result, the Financial Industry Regulatory Authority (FINRA), which oversees the securities exchanges, reported plans to reconsider the odd-lots policy.

In more recent work, Ye and UIUC colleagues Yao and Jiading Gai, examined effects of increasing trading speed from microseconds to nanoseconds. Their calculations with Gordon and Blacklight, processing 55 days of NASDAQ trading data from 2010, looked at the ratio of orders cancelled to orders executed, finding evidence of a manipulative practice called "quote stuffing" – in which HFT traders place an order only to cancel it within 0.001 seconds or less, with the aim of generating congestion. Their analysis provides justification for regulatory changes, and in September their study was referred to as "ground-breaking" in expert testimony on computerized trading before the U.S. Senate Subcommittee on Securities, Insurance and Investment.

## Beyond Flash Crash: Odd Lots

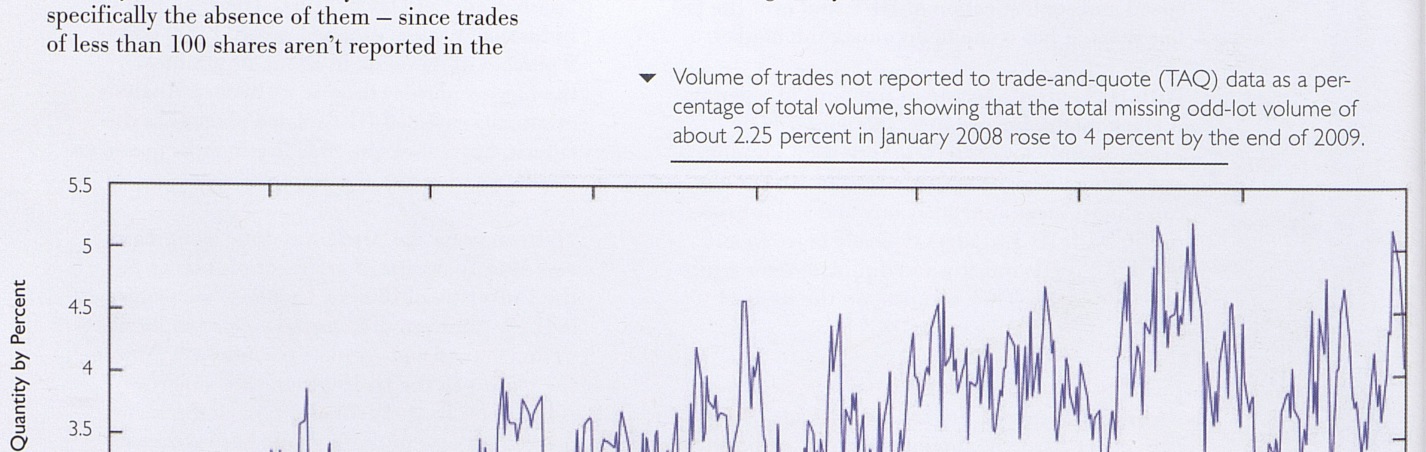
Ye and colleagues recruited Blacklight's shared memory to take on their study of "odd-lot" trades, specifically the absence of them – since trades of less than 100 shares aren't reported in the "consolidated tape" of trade and quote (TAQ) data. The TAQ aggregates trade data across the 13 U.S. stock exchanges as well as several off-exchange trading venues that don't display "bid" and "ask" prices.

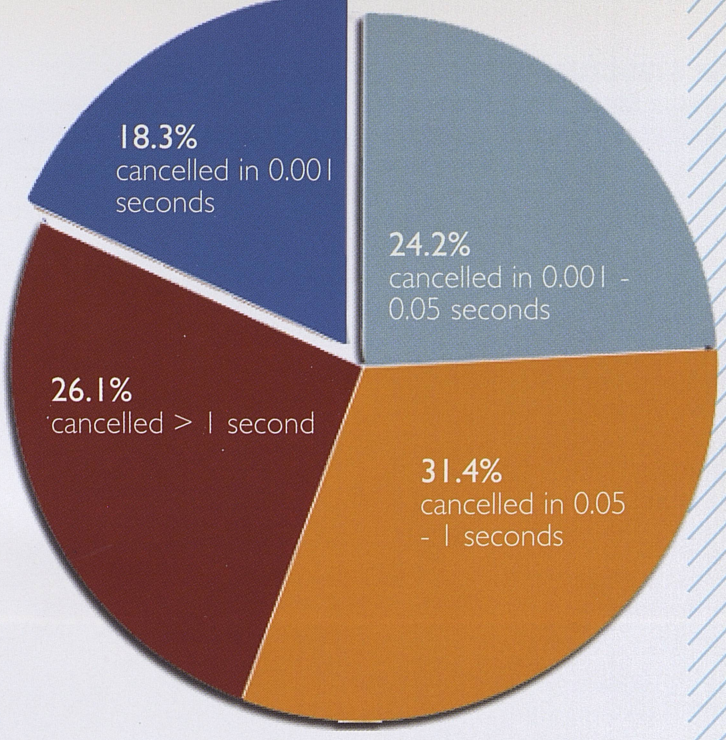
***Because of this XSEDE-supported research, the Financial Industry Regulatory Authority is reconsidering the policy of excluding odd-lot trades from the consolidated tape.***

To assess the implications of the exclusion of odd lots, the researchers relied on two datasets – NASDAQ TotalView-ITCH and NASDAQ high-frequency trading data – that are more comprehensive than TAQ. With Blacklight (and this year augmented by Gordon), the researchers analyzed a large cross-section (7000 stocks) and time series of data: two years – January 2008 to January 2010. For the TotalView-ITCH data, Blacklight's shared memory could store all the files, a total of 7.5 terabytes, at one time – saving considerable time, says Ye, for some of the analyses. Ye accessed the NASDAQ high frequency data, about 15 gigabytes, from the research server at the Wharton School of Business.

In a series of computations, the researchers compared their comprehensive data with the commonly used (but incomplete) TAQ. They found that, due to HFT, odd-lot trades increased from 2.25 percent of volume in January 2009 to 4 percent by the end of 2009. The median number of missing trades per stock was 19 percent, while for some stocks missing trades are as high as 66 percent of total transactions. For the two-year period they studied, they found, furthermore, that 30 percent of "price discovery" – the amount of price change during a day of trading – was due to odd-lot trades. "This is huge," says Ye.

▼ Volume of trades not reported to trade-and-quote (TAQ) data as a percentage of total volume, showing that the total missing odd-lot volume of about 2.25 percent in January 2008 rose to 4 percent by the end of 2009.






**Fleeting Orders ▸**
On August 30, 2011, about three-million orders were submitted to the NASDAQ exchange to trade the stock SPDR S&P 500 Trust (ticker symbol SPY). This image shows that 18.3 percent of the orders were cancelled within one millisecond, and 42.5 percent of orders had a lifespan of less than 50 milliseconds, less time than it takes to transfer a signal between New York and California. More than 40 percent of orders, in other words, disappeared before a trader in California could react.

Many odd-lot trades, their analysis showed, are the result of informed traders splitting orders. Suppose you want to trade 10,000 shares, but you slice it – through HFT – into 200 trades of 50 shares. "If you trade in large lots," explains Ye, "people will guess something has happened and they can follow you. If you trade quietly through slicing into small lots, it looks to other people like no trade has happened."

Prior to HFT, odd lots were a much smaller fraction of market activity, less than 1 percent of New York Stock Exchange volume in the 1990s, and their omission from TAQ wasn't of major consequence. "Because odd-lot trades are more likely to arise from high-frequency traders," the researchers write in their paper, "we argue that their exclusion from TAQ raises important regulatory issues."

Ye and colleagues' findings aroused discussion and were reported, among other places, in *Business Week* and *Bloomberg Businessweek*. Motivated by their study, the Consolidated Tape Association, a group of stock exchange executives that administers price and quote reporting, appointed a subcommittee to look at the implications of the truncated odd-lot data. *Bloomberg* reported in September that FINRA planned to vote in November on whether to include odd-lot trades in the consolidated tape.

## Fleeting Orders & Quote Stuffing

Relying again on the NASDAQ TotalView-ITCH data, Ye, Gai and Yao this year looked at the effects on the market of increasing the speed of trading from microseconds to nanoseconds. For this work, which mainly used SDSC's Gordon but also did some of the analysis on Blacklight, the researchers analyzed "fleeting orders" – orders that are canceled within 50 milliseconds of being placed.

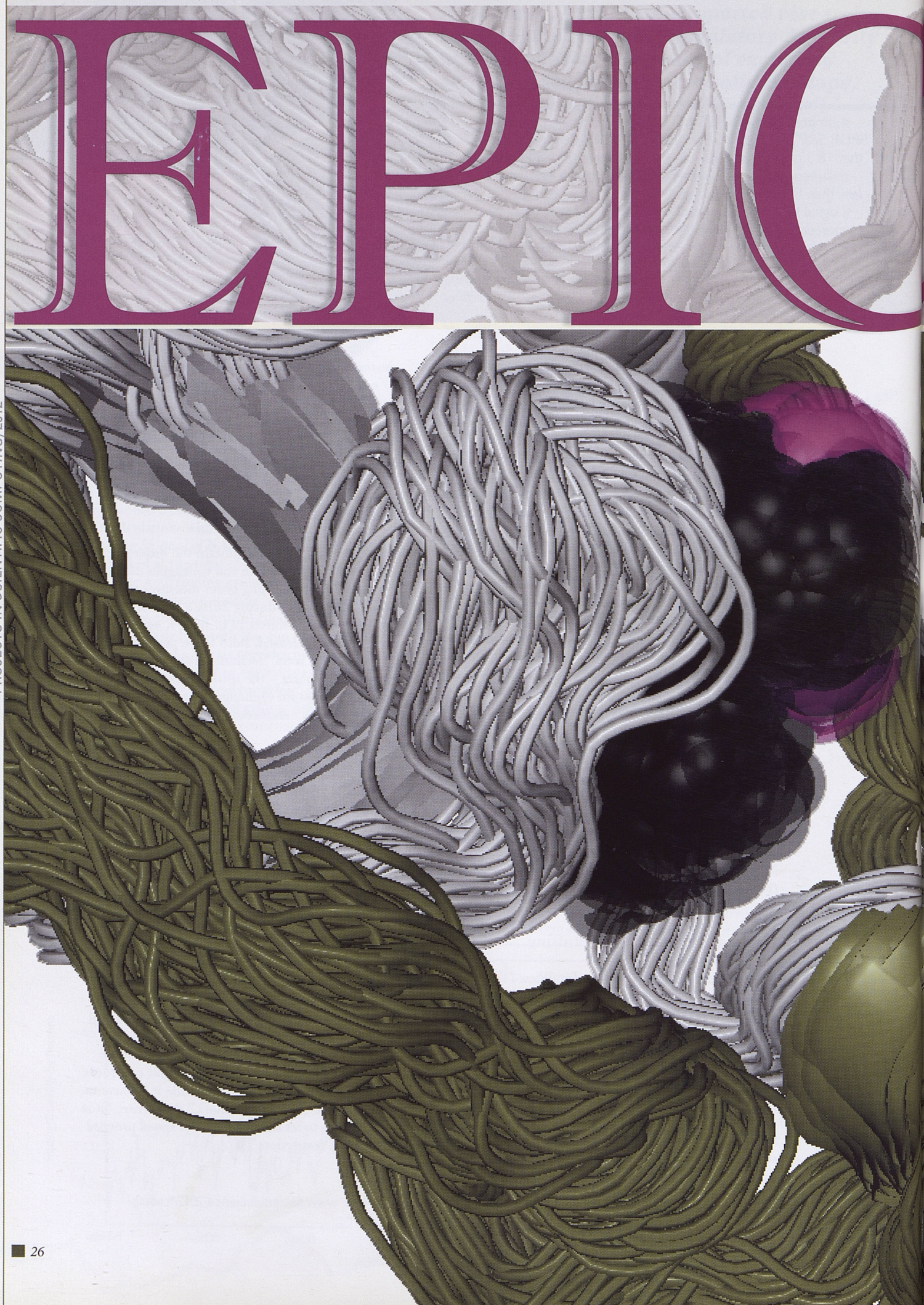
Studies of the May 2010 "flash crash" have suggested that HFT speed in executing and canceling orders may have contributed to the sudden price drop. As a result, proposals for regulation have suggested a minimum quote life or a cancellation fee, which could be based on the average number of order cancellations to transactions.

Processing data files that contain the order instructions for stocks, Ye and colleagues did an "event study" – analyzing order messages that covered two periods during 2010, a total of 55 trading days from March 19 to June 7, when trading speed rapidly increased. Both these periods, which the researchers term "technology shocks," occurred on weekends, when – the researchers note – "it is more convenient for exchanges or traders to test their technology enhancement."

Their paper, the researchers write, is "the first paper to explore the impact of high frequency trading in a nanosecond environment." They found that as trading frequency increased from microseconds to nanoseconds the order cancellation/execution ratio increased dramatically from 26:1 to 32:1. Their analysis found no impact on liquidity, price efficiency and trading volume, but found evidence consistent with quote stuffing – a high volume of trade aimed at congesting the market.

The increase in speed from seconds to milliseconds, say the researchers, may have social benefit by creating new trading opportunities, but they doubt whether such benefit will continue as speed goes from micro to nanoseconds, or possibly, to picoseconds. Their analysis gives justification for regulatory changes, such as a speed limit on orders or a fee for order cancellation. "While it is naive to eliminate high frequency traders," they write, "it is equally naive to let the arms race of speed proceed without any restriction." ■

**More info:** www.psc.edu/science/2012/trading/


EPIC

# Microseconds

## Studies with Anton, a special-purpose supercomputer designed by D. E. Shaw Research and made available to the research community through PSC, have yielded invaluable insights into the motion and function of proteins

Proteins are the action heroes of the body. As enzymes, they make reactions go a million times faster than they otherwise would. As transport vehicles, they carry oxygen, ions and neurotransmitters. Proteins in the cellular membrane act as gated channels to allow water, ions or other biomolecules to pass into or out of the cell. Other proteins within the cell participate in complex biochemical cascades of interaction that keep the cell's internal machinery humming.

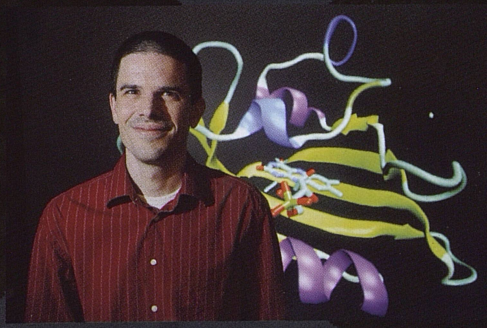
The ability of proteins to do their life-sustaining work, whatever the task, depends on their three-dimensional shape. The helices, sheets and turns of a fully formed protein create hooks and crevices — distinctive features of structure, electronic charge and other properties — that allow proteins to bind and interact with other biomolecules. In these interactions, life happens.

Similarly, when proteins malfunction it's often due to a structural deformation, and many disease conditions are related to malfunctioning proteins. Many drug therapies work by interacting with structural features of proteins in ways that block the malfunctions that cause disease. Rational "design" of more effective drug therapies depends, first of all, on precise understanding of the target structures.

All this is to say that a major task of molecular biology, beginning even before Watson and Crick unraveled the mystery of DNA's double helix, has been bringing to light the three-dimensional structure of proteins, each one distinct from others. It can be years of work to solve one structure, with x-ray crystallography or NMR spectroscopy as the primary tools — along with physics and the insight of experienced researchers. Many thousands of protein structures have been solved this way, and many more thousands remain to be solved.

It's a great scientific accomplishment, yet those solved structures are only a first step. They are static representations, and proteins never rest. As they carry out their activities in the body, they are constantly shifting shape, a sub-microscopic dance of life, and these minute changes in structure can account for the difference between health and disease.

Researchers can "see" these movements in great detail on a computer — by an approach called molecular dynamics (MD). Starting with a protein's static structure, MD models the forces that act between atoms and can thereby track the precise, atom-by-atom details of a protein's motion. At least for a few nanoseconds . . . a start, but only that.



▲ **Markus Dittrich,** PSC, NRBSC. "We're thrilled about the impact that Anton has had over the last two years," says Dittrich, "and we're excited to offer continued access to this great resource."

Enter Anton: a supercomputer designed and built – by D. E. Shaw Research (DESRES) in New York City – to dramatically increase the speed of MD. Named in honor of Dutch microscope inventor Anton van Leeuwenhoek, Anton allows researchers to see previously unseen biomolecular activity. By executing ultra-fast MD, Anton simulates proteins (and nucleic acids) for longer stretches of biological time than was previously possible. Before Anton, even the most powerful supercomputers could – because of the prohibitive amounts of computing required – track a protein's movement for only hundreds of nanoseconds ($10^{-9}$ seconds), with a few MD simulations reaching into the microsecond range ($10^{-6}$ seconds).

"Anton performs MD simulations up to 100 times faster than conventional supercomputers," says Markus Dittrich of PSC's National Resource for Biomedical Supercomputing (NRBSC), who directs the Anton program at PSC, "making it possible for the first time to simulate the behavior of proteins over more than a millisecond of biological time. The availability of these extended timescales has opened a new window on many important biological processes."

Thanks to DESRES, which provided a machine to NRBSC without cost, an Anton system has been available at PSC since late 2010 for use by the general biomedical community. A two-year $2.7 million grant to NRBSC from NIH's National Institute of General Medical Sciences provided initial support for operational costs. To date, 70 research groups have used Anton at PSC for work on 91 projects, with a new round of allocations underway. This work, including the four projects described in the rest of this article, has led to many new findings.

## The Protein-Folding Problem: A Big Step

A droopy, strung-out chain of amino acids – that's what rolls off the assembly line of the molecular factory inside a cell when a protein is created. All the pieces are there, and they're in the right sequence. But the new protein is unfit for duty.

To do its job, this dangly chain must fold into just the right three-dimensional configuration. It happens within seconds or less, and the result is a complex bundle of helices, sheets and turns ready-made for the protein to go to work.

How does this happen? Out of billions of possible shapes that the chain of amino acids could assume, how does it arrive at the shape it takes in nature? "It's a famous problem," says biophysicist Klaus Schulten of the University of Illinois, Urbana-Champaign, "and only recently, because of computing, has it been coming close to solution. Experiments measure too few specifics of this process to make a statement. Only the computer, with MD simulations, can follow the detailed processes involved when strands of amino acid arrange themselves into a protein in proper form."

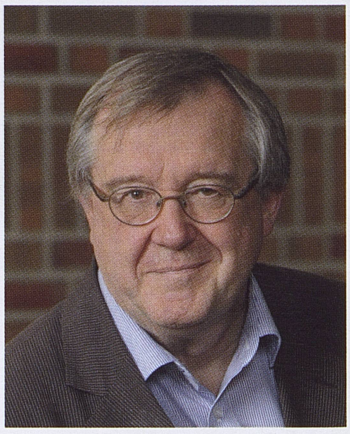
***"This field is undergoing a revolution. With Anton, we're able to fold larger, more natural proteins."***

▲ The native state, final folded structure of lambda-repressor, a five-helix bundle, colored blue to red from the N-terminus (the amine group, by convention the starting end of a protein's structure) to the C-terminus (carboxyl group).

Using Anton, Schulten – teamed with experimentalist Martin Gruebele and UIUC colleagues – successfully simulated the folding of an 80 amino-acid protein (lambda-repressor). With water molecules and ions, the simulation included 74,253 atoms. Their findings (*Journal of Physical Chemistry Letters*, April 2012) showed a folded result in good agreement with experiment, and went beyond experiment to find that accepted ways of measuring folding in the laboratory give an incomplete picture. Further experimental work, and a follow-up paper, are underway.

"This field is undergoing a revolution" says Schulten. "It began with folding very small proteins, but now with Anton, we're able to fold larger, more natural proteins. Lambda-repressor is one of the largest proteins the folding of which has been monitored and described in the computer. It's a stepping stone toward solving the very important protein-folding problem."

The researchers in total tracked 100 microseconds of protein movement, 80 microseconds with Anton – in two separate 40 microsecond simulations, each of which took a week. The additional 20 microseconds with another computer took a year, explains Schulten: "We could do this only with Anton."



▲ **Klaus Schulten,** University of Illinois, Urbana-Champaign

## Stop the Bleeding: Inside-Out Signaling

Integrins are the essential two-way communicator proteins of cellular biology. This large family of receptors, proteins that reside in the cellular membrane, receive information about things happening in the cell's external environment, the extracellular matrix (ECM), which triggers a cellular response. For example, integrins marshal the body's response to a wound. When exposed to collagen – proteins in connective tissue – at a wound site, integrins on the surface of blood platelet cells change shape, a shift that dramatically increases integrin's binding affinity for fibrinogen, a blood-clot forming protein. Through this process, called thrombosis, fibrinogen binds platelets to each other, a blood clot forms, and – if all goes well – bleeding stops.

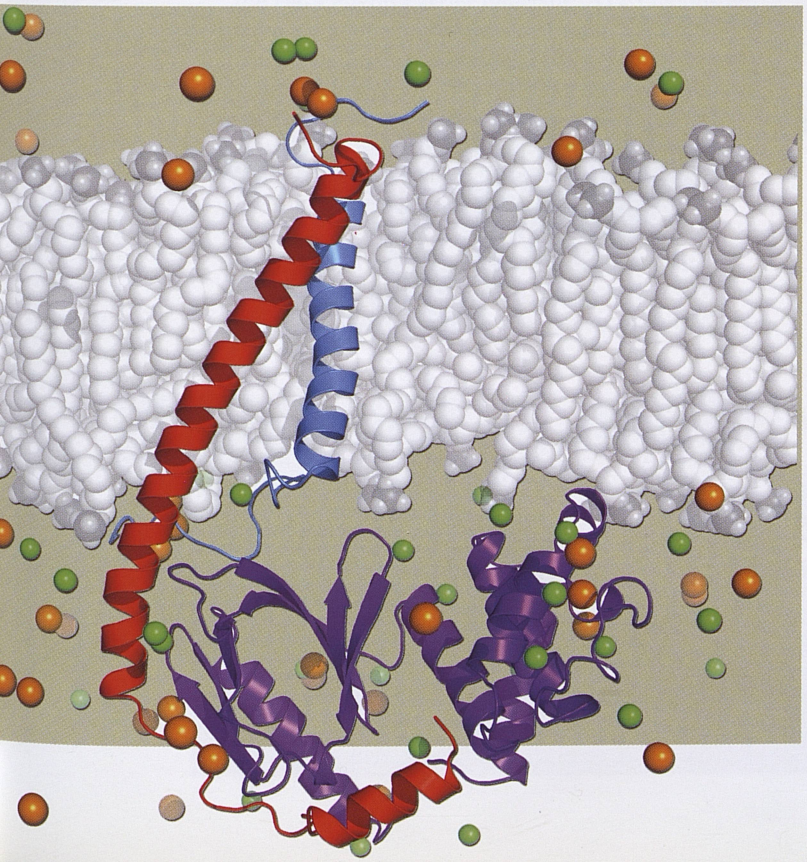
Integrins are also involved in cell migration and immune-system patrolling, among other processes, and along with responding to changes in the ECM, integrins also operate in an "inside-out" mode of signaling. They switch to a different shape – become activated – through interaction with intracellular proteins, one of which is called talin. "Talin binds to the tail of integrin," says Marta Filizola of Mount Sinai School of Medicine, "and it's a hard question to determine how and to what degree this has an impact on activating integrin."

Using Anton, Filizola and colleagues simulated the helical region of the platelet integrin (called αIIbβ3) that lodges in the cellular membrane with a tail extending into the cell. Their MD simulations included a phospholipid bilayer, representing the cellular membrane, surrounding water molecules and ions, along with two talin domains (F2 & F3) – an unprecedented level of detail, about 76,000 atoms. They simulated about five microseconds of talin interacting with integrin and, for contrast, did the same simulation of integrin and the lipid membrane without the presence of talin, about 50,000 atoms. The researchers augmented this second simulation with additional MD simulations on their in-house cluster.

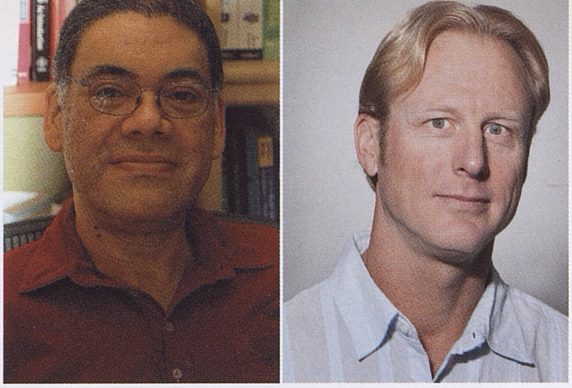
▲ **Marta Filizola** (center) with her lab group at Mount Sinai School of Medicine. The integrin simulations were carried out by Davide Provasi (far left) and Ana Negri (far right).

***They showed atom-by-atom details of the talin-integrin interaction that go beyond previous understanding.***

Their studies confirm a hypothesis that had been put forward by recent experimental work – that parts of talin anchor to the inner membrane wall, helping to stabilize the interaction with integrin. They also showed atom-by-atom details of the talin-integrin interaction that go beyond previous understanding and suggest directions for more experimental studies. Specifically, they showed, says Filizola, that talin reduces a tilted orientation of integrin's two helical subunits, and induces bending of integrin's β3 helix: "These simulations, which we couldn't have done without Anton, broaden our understanding of how talin contributes to integrin's activation."



◀ A snapshot of the end point of a five-microsecond MD simulation by Marta Filizola and colleagues of integrin interacting with talin. The two transmembrane helical subunits of integrin — *αIIb* (blue) and *β3* (red) — are shown with the tail of the *β3* helix that extends into the cell interacting with the F3 domain of talin (purple). Lipid head groups (gray spheres) form the boundaries of the cellular membrane, and surrounding ions (colored spheres) are also included. For clarity, the lipid chains that form the front part of the lipid membrane and water molecules are omitted.



▲ Alfredo Freites (left) & Doug Tobias, University of California, Irvine

## Opening the Gate to Action

When the starting gun sounds and Usain Bolt explodes from the blocks in the 100-meter dash, electrical signals in the brain trigger his blastoff. Voltage-gated ion channels open in nerve cells, and ions – charged atoms, usually sodium and potassium – flow through the opened gates, creating electrical currents that cause muscle fibers to contract.

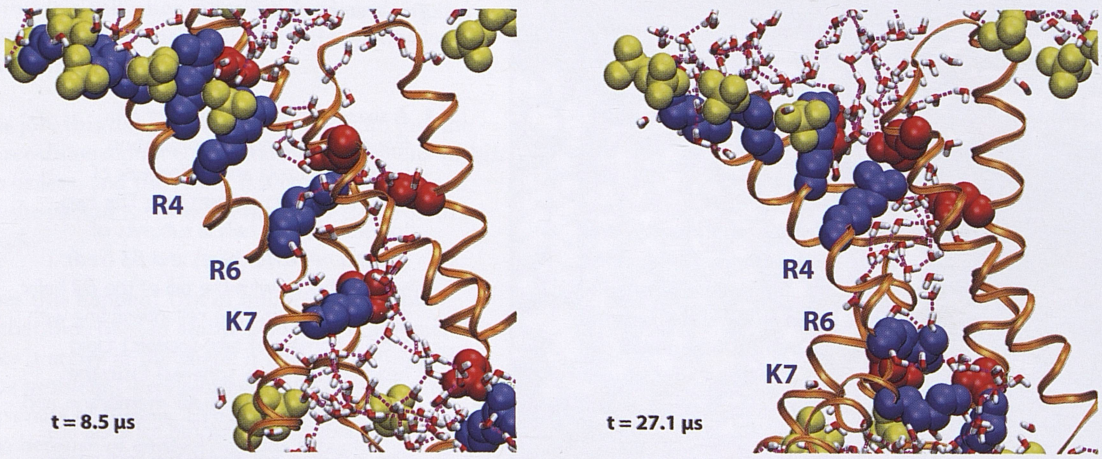
"Every communication in the central nervous system is possible because ions flow across the cell membrane," says Alfredo Freites, a biophysicist with the group of Douglas Tobias at the University of California, Irvine. The ions flow, he explains, through what's essentially a hole in the membrane formed by proteins, called voltage-gated ion channels. These channels open and close based on the ability of part of the protein, called the "voltage sensing domain" (VSD), to respond to changes in electrical potential.

Laboratory studies over many years have shown that currents – called "gating currents" – in the VSDs are associated with motions that trigger opening of the channel. Until the availability of Anton, however, it hadn't been possible to track what structural changes happen in the VSD during the gating event. "Despite a wide variety of data," says Freites, "the molecular mechanism of voltage gating hasn't been well understood."

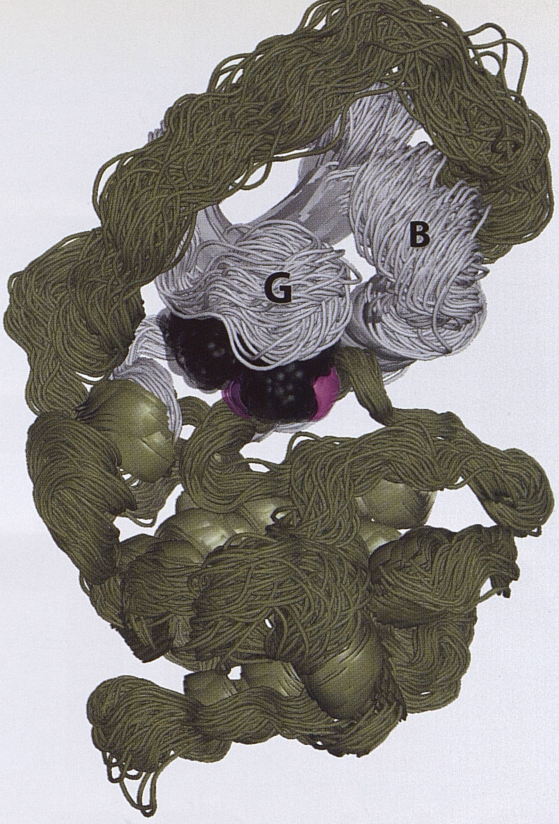
With Anton, Freites and his colleagues Eric Schow, Stephen White and Douglas Tobias were able to run MD simulations over a timescale that corresponds to a VSD gating-current event. They simulated the VSD embedded in a lipid bilayer, representing the cellular membrane, along with surrounding water with an applied electric field for a period of 30 microseconds. At this timescale, the researchers were able to make direct comparisons between MD simulations and laboratory data. "With any other high-performance computing resource," says Freites, "it would be impractical to do this."

***"With any other high-performance computing resource, it would be impractical to do this."***

Their findings – reported in *The Biophysical Journal* (June 2012) – were, in general, consistent with the data from laboratory studies. The detail of the MD results suggests, nevertheless, that gating-charge measurements from electrophysiological lab studies "may not represent a single charge displacement but may instead be the superposition of many events occurring faster than the instrument response." Their findings also go beyond prior studies, observes Freites, in showing that the presence of water molecules within the VSD is necessary for the gating current to flow and pull the channel open. "Water facilitates the flow of charges," says Freites, "and we see that the VSD's own hydration is what allows this event to happen seamlessly."

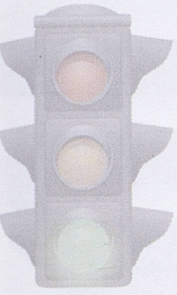



▲ **Before & After**
Snapshots of the VSD (orange) with associated lipid bilayer phosphate groups (yellow) before (left) and after a sudden change in the electrical potential at nine microseconds. The displacement of three highly-conserved, positively charged amino-acid residues (blue — R4, R6 & K7) gives rise to a gating-current event. As these charges move through the water molecules (white and red) in the VSD interior, they exchange interactions with negatively charged amino-acid residues (red).




**Motion at PKA's Active Site** ▸
This graphic shows overlaid snapshots from one of the simulations by McClendon and Taylor. The two main structural components of PKA's catalytic domain, the N-lobe (light gray) and C-lobe (dark green), enclose the "active site," which holds ATP (black) and two magnesium ions (purple). The image, observes McClendon, shows that PKA's glycine-rich loop (G) and nearby B-helix (B) "aren't locked down by the ATP and two magnesiums, but instead remain more flexible than we expected."

## Action at the Crossroads



Think of a traffic cop at a crazy downtown five-way (or more) intersection. The officer in blue is performing a graceful dance – waving cars through from one direction, holding them off from another, switching from stop to go, go to stop, vehicles and their payloads of goods and busy people on intersecting paths getting to where they need to far more smoothly and expeditiously than you thought possible.

In a cell of the human body, one such traffic cop is protein kinase A (PKA) – a crude analogy, but one that roughly conveys the role played by this protein in regulating the complex network of chemical reactions within the cell. PKA is one among a superfamily of enzymes, the kinases, that are ubiquitous in living things. "Protein kinases operate like stop and go signals," says Susan Taylor, professor of chemistry and biochemistry at the University of California, San Diego. "They are essential molecular switches for all biology."

PKA is the prototype of this big family, a protein for which Taylor and her colleagues first solved the structure in 1991. With this structure as a map, Taylor's research group has answered many questions about how protein kinases regulate cell metabolism by means of a biochemical handoff called "phosphorylation." Protein kinases take the cell's energy-carrier molecule, ATP (adenosine triphosphate), and transfer a phosphate group from it to target proteins, often altering the target proteins' functions. Having given up a phosphate, the ATP becomes ADP (adenosine diphosphate). Kinases then let go of the ADP and become active again once they bind with another ATP molecule and find another target protein to phosphorylate.

By phosphorylating a variety of target proteins, PKA helps regulate memory, cell growth and many other processes. When kinases go awry, diseases are often the result – especially cancer. For this reason, the kinase superfamily, with PKA as prototype, is a prominent target for drug therapy, and several effective anti-cancer drugs that work by blocking the active site of defective kinases are already available.

***Several anti-cancer drugs that work by blocking the active site of kinases are already available.***

To advance this work, Taylor and post-doctoral researcher Chris McClendon used Anton to simulate several different states of PKA's catalytic domain, which binds with ATP and releases ADP. NMR studies by Gianluigi Veglia at the University of Minnesota, in collaboration with the Taylor lab, showed that these processes occur on slow biological timescales of milliseconds. With availability of Anton, McClendon was, for the first time, able to glean useful information from MD simulations about these cyclical structural changes.

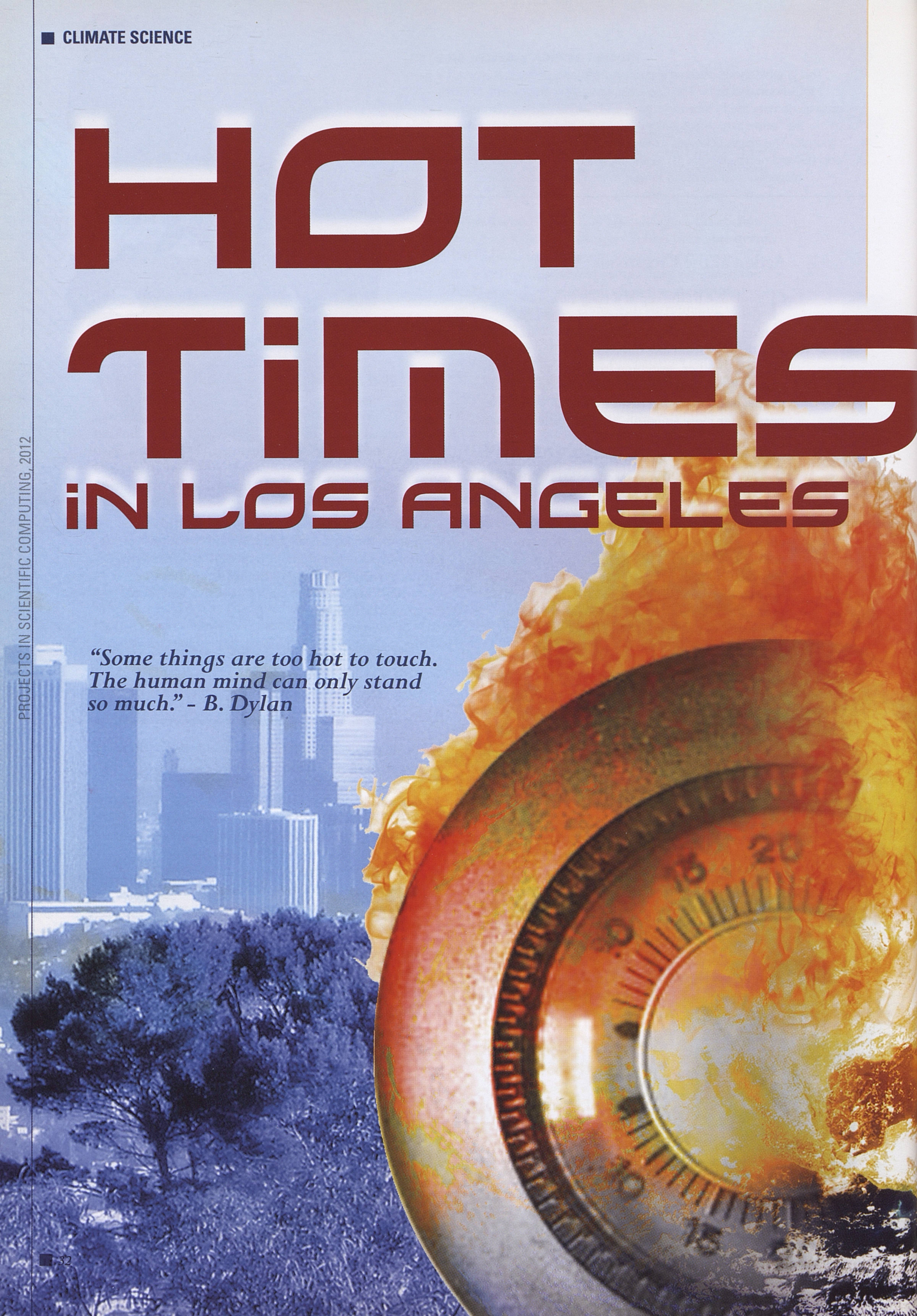
A key finding was that the tail of the C-lobe, at the top of the PKA structure, flips open and closed like a latch to hold and release a "gylcine-rich loop" that closes over the ATP. The simulations also suggest that active-site opening and closing can occur at rates faster than expected from prior studies. "We're getting new clues as to what regions are dynamic," says Taylor. "It's the first time we can do calculations at a timescale we can experimentally validate. Anton provides us with a way to test the consequences of disease mutations and engineered mutations on the overall dynamics, which we have never been able to do." ■

**More info:** www.psc.edu/science/2012/antonepics/

▲ Susan Taylor & Chris McClendon, University of California, San Diego

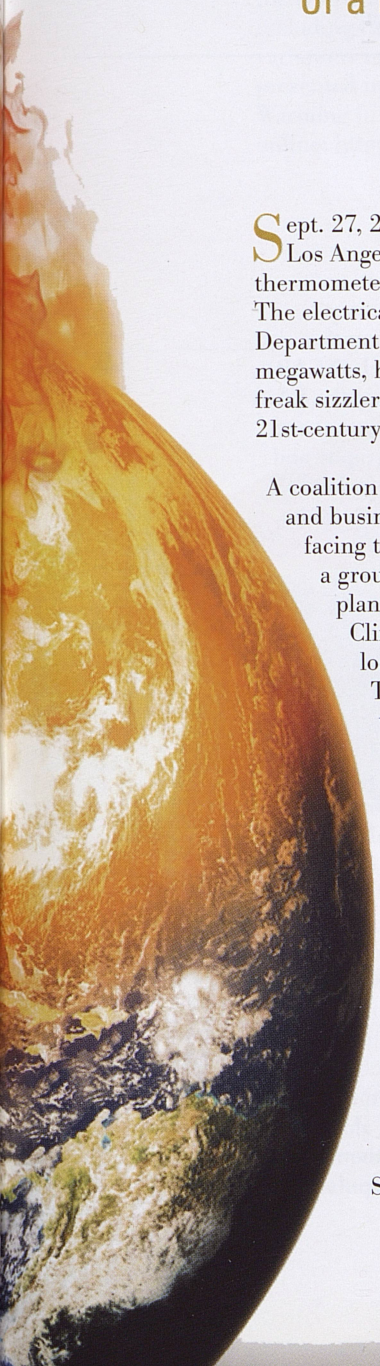
# HOT TIMES IN LOS ANGELES

*"Some things are too hot to touch. The human mind can only stand so much." - B. Dylan*



**Alex Hall**, University of California, Los Angeles. A "very valuable resource," Hall says of Blacklight, PSC's system that supported about half the computations for the LA study. "Because you have a grid where computations are impacting each other, it's very helpful to have shared-memory capability." Hall credits PSC scientist and XSEDE consultant David O'Neal, who supported the LA team. "He was extremely knowledgeable and professional. We have problems sometimes, and to have someone like him easily accessible is very helpful."

## As part of efforts to develop a Climate Action Plan, a Los Angeles area team produced the first study assessing effects of climate change on the scale of a metropolitan region



Sept. 27, 2010: The hottest day on record in Los Angeles. The official weather station thermometer broke when it reached 113° F. The electrical load from the Los Angeles Department of Water and Power peaked at 6,177 megawatts, highest in history. As days go, was it a freak sizzler, or harbinger of the new normal for 21st-century LA?

A coalition of municipalities, academic institutions and businesses in the Los Angeles region are facing this very tough question. Through a ground-breaking initiative in regional planning, they are working to develop a Climate Action Plan that accounts for the local effects of global climate change. To lay a credible foundation for this work, Alex Hall, a professor in UCLA's Department of Atmospheric and Oceanic Sciences, is leading an effort in computational modeling. In June 2012 he and his colleagues released a study, "Mid-Century Warming in the Los Angeles Region," that is the first published study assessing effects of global climate change at the scale of a metropolitan region.

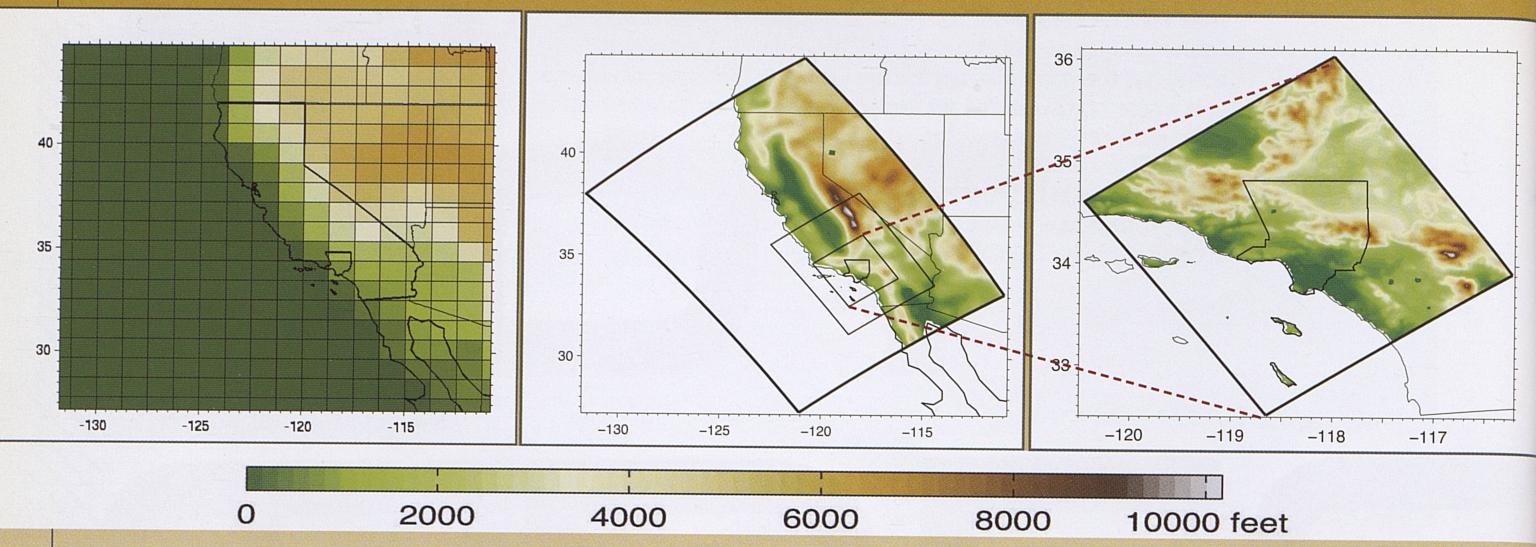
"This is the most sophisticated climate science ever done for a city," said UCLA professor Paul Bunje, who directs the Los Angeles Regional Collaborative for Climate Action and Sustainability.

For computational resources, Hall relied on XSEDE – in particular, PSC's Blacklight (and, initially, Ember at NCSA) along with the National Energy Research Scientific Computing Center (in Berkeley, California) and UCLA in-house computing. The project takes results of global climate models (GCMs) and applies them at the much reduced scale of the greater Los Angeles region. Because GCM results lack the detail needed to give a clear picture at regional scale, Hall and colleagues did extensive calculations to downscale GCM results to the local features.

The study predicts that for the years 2041 to 2060 temperatures in the greater Los Angeles area will be higher, compared to the last 20 years of the 20th century, by an average of 4-5° F. The number of extremely hot days – temperature above 95° F – will triple in the downtown area, says the study, and quadruple in the valleys and at high elevations. "Every season of the year in every part of the county will be warmer," says Hall. "This study lays a quantitative foundation for policy-making to confront climate change in this region. Now that we have real numbers, we can talk about adaptation."

### Zooming In: Dynamical Downscaling

The greater Los Angeles "combined statistical area" – which includes Orange County and parts of Ventura, San Bernardino and Riverside counties – is home to nearly 18 million people. Together they account for nearly $750 billion a year in economic activity. For most of the 20th century, LA was the fastest growing region in the country, due in large part to the Mediterranean-like climate, warm to hot dry summers and mild winters.




**Zooming In on LA**

Even in global climate models with the highest resolution the Los Angeles region is merely a pixel (left). The mountain ranges across the region are completely lost in the global models. The local topography that defines much of the Los Angeles climate is completely wiped away. The grid is laid out by degrees of latitude (vertical axis) and longitude (horizontal axis), with surface altitude represented by the color scale. Downscaling brings the view in closer (center). In the third map, the inner domain of the modeling (with Los Angeles County outlined), what was only a pixel in GCMs has clear topographic detail that defines the varieties of climate.



The area's geography, a coastal basin nested between the Pacific Ocean and Sierra Mountains, is part of the challenge, says Hall, of understanding the local effects of global climate. Among other factors, modeling must account for how these topographical features shape winds, known as the Santa Anas, which fuel some of the most furious wildfires that occur in densely populated areas.

The central challenge of the modeling is the contrast in scale and resolution of GCMs compared to the LA region. GCMs solve the equations of the atmosphere – wind, clouds, surface temperature, topography and many other factors – over a computational grid (roughly 100 kilometers on a side) that covers the world. Computing power, even at current petascale levels, isn't enough for GCMs to include detailed topography of each urban region. "Even in global climate models with the highest resolution," says Hall, "the Los Angeles region is merely a pixel."

To bridge from the scale of GCMs to a metropolitan area – to reliably capture information at fine enough detail on which to base planning, Hall and colleagues applied an innovative two-stage approach that drew on the archived results of 19 GCMs. "These global models are done on supercomputers around the world," says Hall, "and the output is publicly accessible."

The first stage was a demanding computation called "dynamical downscaling." For this step, the researchers used a regional model from the National Center for Atmospheric Research (NCAR) as their software framework. They overlaid the LA region with a fine grid (two kilometers square) and initialized the model with regional topography and actual atmospheric data (called "reanalysis data"), including ocean-surface data, at coarse resolution from archives.

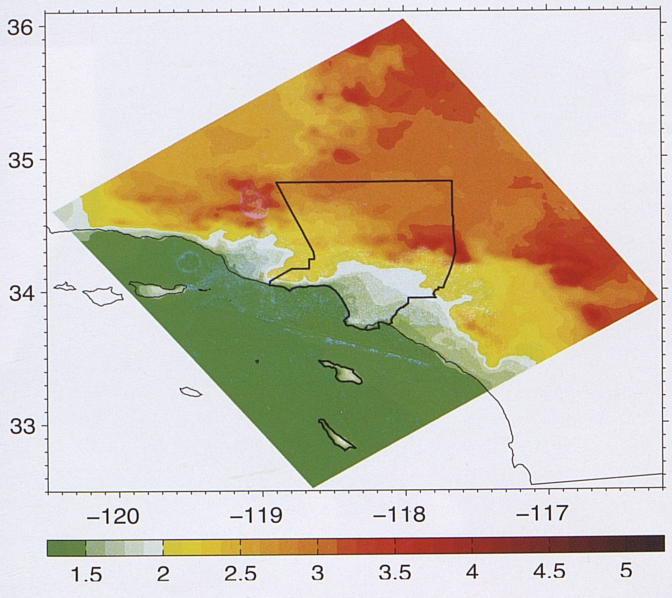
The first dynamically downscaled simulation established the "baseline," reconstructing LA regional weather (rainfall, surface temperatures, wind directions and speed, etc.) at fine detail for 1981-2000. Results from this simulation compared well with historical data from 23 weather observation sites in the LA region, lending confidence to the approach. "We can reproduce rain events," says Hall, "when they actually occurred going back as far as data is available."

With 1981-2000 as their validated baseline, the researchers then ran the model again. Drawing on output from GCMs, including an NCAR GCM for North America, this second large simulation generated a forecast of the LA regional climate for the 20-year mid-century span of 2041-2060. Each of the two dynamical-downscaling simulations used 96 of Blacklight's processors (six blades), along with UCLA in-house computing – altogether about four months for each of these demanding computations.

***Even if the world succeeds beyond expectations in cutting back greenhouse emissions, says the model, the LA area will still get 70 percent hotter.***

"The main advantage of dynamical downscaling," says Hall, "is that the regional numerical model produces a climate change response driven purely by its own internal dynamics. It is in no way predetermined by any assumptions about the relationship between regional climate and climate at larger scales."

In the second stage of their strategy, the researchers applied statistical techniques, far less computationally intense than dynamical downscaling, to incorporate results from other GCMs into their mid-century forecast. This technique used parameters (derived from the dynamically downscaled baseline modeling) to represent patterns in the relationship between the LA region and the NCAR GCM, in this way including the benefit of a wide range of GCM variations in the LA results. The overall outcome, presented in June as a white paper, "Mid-Century Warming in the Los Angeles Region," is available online: http://c-change.la/pdf/LARC-web.pdf.




▲ **Surface Warming**
This graphic shows the change in warming (difference between the 1981-2000 baseline and the 2041-2060) as an annual mean surface air temperature in °F, increasing from green to red. "Note the contrast," says Hall, "between inland and coastal warming, and stronger warming at higher elevations."

## Hotter Days & More of Them

The 2041-2060 model presented two scenarios for LA conditions at mid-century – "business-as-usual" with greenhouse gas emissions continuing, and a scenario with reduced emissions. The model shows that even if the world succeeds beyond expectations in drastically cutting back greenhouse emissions, the greater LA area will still warm to about 70 percent of the business-as-usual scenario. "I was a little taken aback," says Hall, "by how much warming remains, no matter how aggressively we cut back."

The white paper reports that the number of hot days, when temperature climbs above 95° F, will increase two to four times, depending on location. Temperatures that now occur only on the seven hottest days of the year will happen two to six times as often. The model detail facilitates neighborhood-by-neighborhood findings – showing, for instance, that coastal areas like Santa Monica receive less warming than inland.

While yearly average temperature increases, the most intense effect is in the hot months, which get hotter – with less warming in spring and winter. The modeling shows also that the inland mountains experience increases in average temperature similar to desert areas – an average increase above baseline of about 5° F.

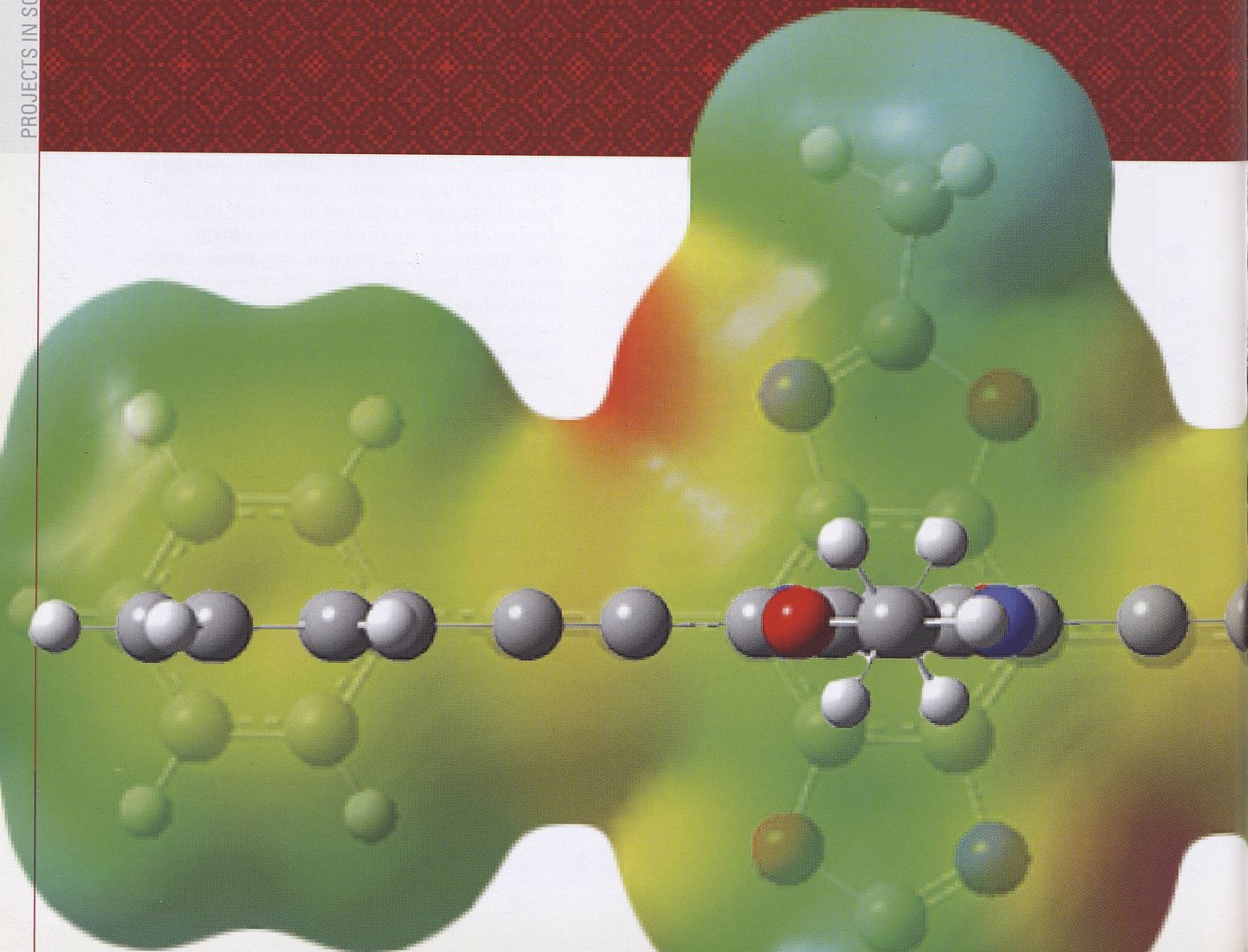
"We've provided some matter-of-fact information about future conditions," says Hall. "It's not meant to be alarming, but to turn this into a problem to be addressed." Along with a California statewide effort to reduce greenhouse gas emissions, several regional adaptation programs are already in place, observes Hall, such as increasing tree canopy – including an active "green roof" movement – and programs to build more parks and open space. The neighborhood-scale projections of Hall's "Mid-Century Warming" study have also given impetus to a plan for a network of air-conditioned heat trauma centers.

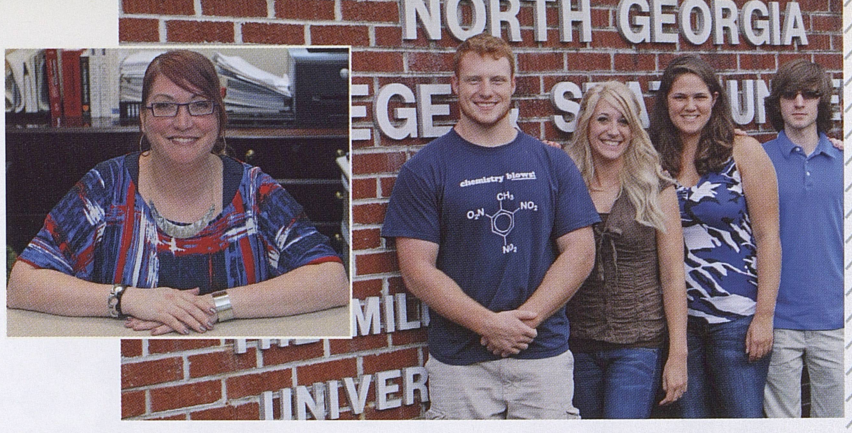
In further work, the LA modelers are collaborating with fire experts to look in detail at the implications for wildfire from the Santa Ana winds. They are also investigating the LA "June gloom" phenomenon, when in May and June, notes Hall, the marine layer spreads inland and cloudiness takes over the region. "We see that this is impacted by changing climate." The modeling team is also delving into key questions, such as the effects of snowpack and low clouds, that affect the availability of water. "With supercomputing," says Hall, "we can simulate these phenomena in detail and see why they change and assess the credibility of these changes." ■

**More info:** www.psc.edu/science/2012/climate/

# CONJUGATE YOUR POLYMERS

Using Blacklight and other XSEDE resources, a computational chemist shows the feasibility of "tuning" carbon-based semiconductors, thereby mapping a path around the roadblocks of silicon-based electronics






▲ **Aimée Tomlinson** (inset) has trained undergraduates at North Georgia College & State University in using XSEDE resources that include Abe, Cobalt and Ember at NCSA and Trestles at SDSC along with Pople and Blacklight at PSC. This work has produced several student posters that have been presented at local, regional and national meetings.

Our ancestors had stone, bronze and iron. For us, silicon defines our age. This material's unique properties as a semiconductor make it the foundation of modern electronics, and over a half century of innovation in smaller, faster circuitry – with silicon as the base material – has transformed the way we live. One of today's smart phones has more computational power than Apollo 11. Our businesses run on computers and Internet trade. Our military conducts war via satellites and computerized drones. It's not an exaggeration to say that America's economic and physical security depend on silicon.

While ever more condensed circuitry – better performing silicon "chips"– has led us to where we are, roadblocks loom. Gordon Moore, for instance, of "Moore's Law" – which holds that silicon devices will double their circuit density every two years – has predicted that this progress will eventually stall, as miniaturization is inherently limited by the size of atoms.

Awareness of these limitations drives research seeking workarounds to silicon. One of the most promising may be carbon – in the form of "conjugated polymers," organic long-chain molecules that conduct electricity and can act as semiconductors. The 2000 Nobel prize in chemistry recognized three chemists – Alan Heeger, Alan MacDiarmid and Hideki Shirakawa – for the discovery and development of these molecule-sized wires.

"Conjugated polymers," says Aimée Tomlinson, "have the potential to revolutionize the world of semiconductors." Tomlinson is a computational chemist at North Georgia College & State University. She used Blacklight at PSC and other XSEDE resources in Illinois (NCSA) and the Trestles system at San Diego Supercomputer Center to study materials, called benzobisoxazoles (BBOs), that can be used as chemical building blocks to produce conjugated polymers (CPs). Her work is aimed particularly at solar cells.

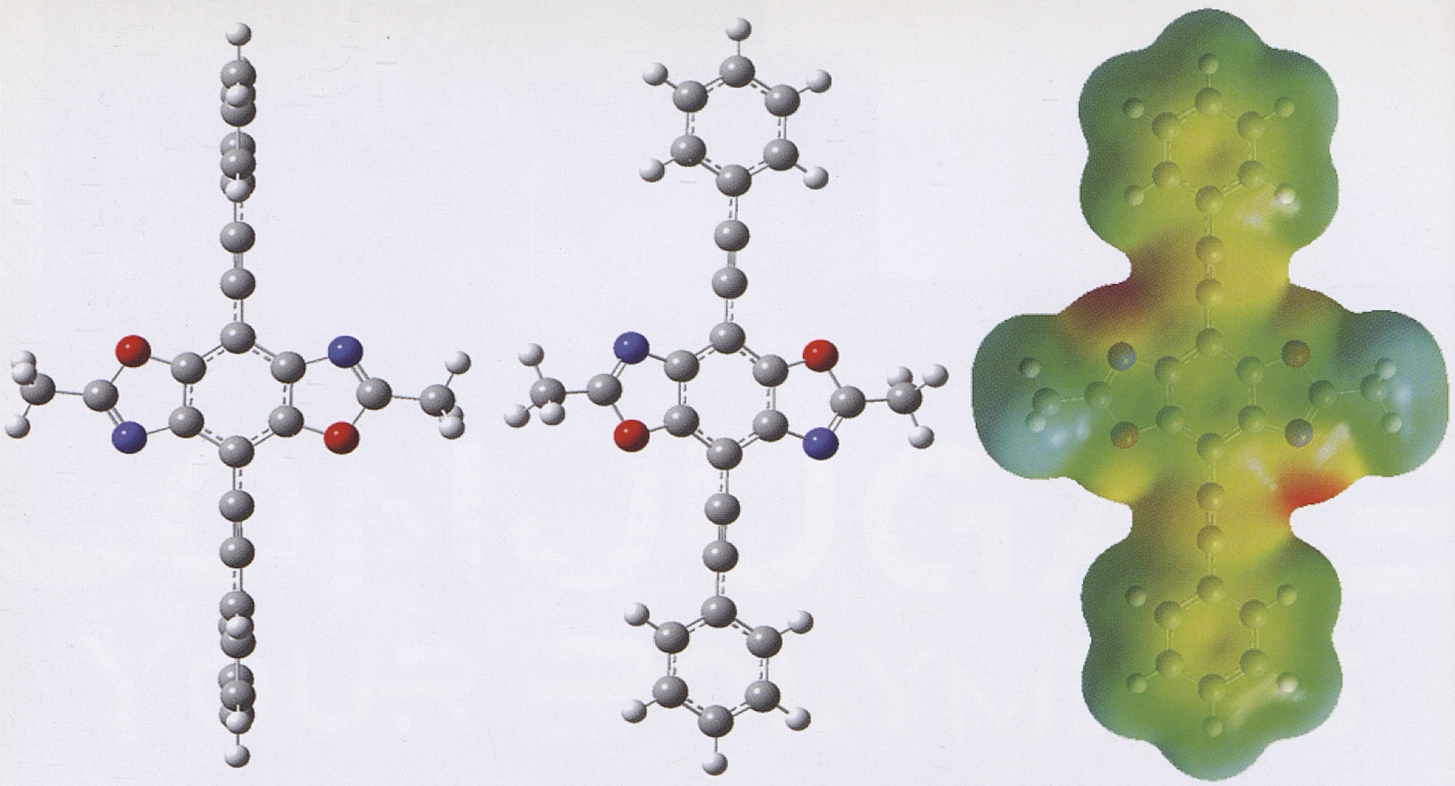
"Organic materials provide several benefits over inorganics for solar cells," says Tomlinson. They are "greener" to mass produce, less detrimental to the environment, she points out, and they have plastic-like properties that are advantageous in industrial processing, since they can be more easily molded than inorganics like silicon into various shapes.

Tomlinson worked with a team of her undergraduate students and collaborated with synthetic chemist colleagues at Iowa State University. Her quantum computations predicted behavior that closely matched experimental findings from lab-synthesized BBOs. Their results – reported in *The Journal of Organic Chemistry* (October 2011) – show that electronic properties of CPs can be varied in predictable ways, saving trial-and-error lab work, to reliably forecast desired electronic behavior, work that reinforces the promise of CPs.

## Tuning the Orbitals

Compared to inorganic semiconductors, CPs are cheaper to manufacture, observes Tomlinson. Along with their plasticity, they are also greener – the production process doesn't rely on toxic metallic inorganics. And in theory they offer a far greater versatility in electronic behavior.

In much the same way that silicon can be "doped" – by the addition of impurities – to produce specific electronic properties, CP organic "wires" can be doped by adding molecular side groups along the main chain. "From a theoretical point of view," says Tomlinson, "if you add a chemical group




▲ **A Simple Twist**
For this BBO compound, one of several that Tomlinson modeled, her calculations showed that the 90° twisted geometry (left) gave closer correlations with experimental findings than the planar geometry. "Nonplanarity is supported by other research," explains Tomlinson, "and we know that the experiments are in solution where these compounds have flexibility to move around." The associated "electrostatic potential map" (right) shows polarity of charge, positive (blue) to negative (red). "The red region is electron rich and becomes increasingly electron poor as we go from red to blue."

that tends to donate electrons, that will make your polymer chain more of an electron donor, willing to transfer electrons to other molecules. If you add a group that withdraws electrons, that's going to be more of an electron accepting material."

But you need to "tune" a donor and an acceptor for them to work properly with each other, she adds. This requires knowing how side groups will affect the critical parameters – the "highest occupied molecular orbital," or HOMO, and the "lowest unoccupied molecular orbital," or LUMO. HOMO describes the highest-energy electrons in a polymer at rest; LUMO describes the higher energy level needed to start them moving in an electrical current. They're both important for telling whether a polymer is likely to work for a particular device.

To capture light energy in a solar cell, which means to convert photons – quantum particles of light – into electricity, the energy of the photon, which is determined by its color (the wavelength of light involved) has to equal the difference between the HOMO and LUMO. For a pair of molecules, the energy difference between the HOMO of one and the LUMO of the other must be just right for the electrons to make the jump.

Tomlinson used the quantum chemistry software GAUSSIAN09 to compute the electronic properties of BBOs. Carnegie Mellon chemist John Pople received the 1998 Nobel prize in chemistry for his work in the development of GAUSSIAN and its underlying theory, which provides a way – with the computational power of modern high-speed processors – to solve the quantum equations that govern electrons and provide accurate predictions of their behavior.

Blacklight's shared memory was essential for these GAUSSIAN calculations, says Tomlinson, because the volume of data involved in the computation goes up with the square of the number of atomic orbitals in each atom, multiplying between atoms. "The shared memory helped tremendously," she says. "The more memory you can put on these calculations, the faster they go."

***The ability to fine tune electronic properties by changing side groups may allow these organic materials to perform tricks that silicon could never match.***

PSC scientist and XSEDE consultant Marcela Madrid helped to coordinate with Tomlinson and notes that her computations made efficient use of Blacklight. "She used approximately 90 percent of the 128 gigabytes of memory available on 16 cores."

The calculations showed deviation between predicted HOMO values for BBOs and those of the synthesized BBOs ranging between 2.4 and 12.8

▲ Marcela Madrid, PSC, XSEDE Extended Collaborative Support Services

percent. Corrections based on more accurate starting molecular structures brought all the errors below 3.5 percent, a reliable range of predictive accuracy to identify compounds to test in a device. The full value of the result, notes Tomlinson, may be best gauged in time. The results give confidence that Tomlinson's computations can predict the HOMO-LUMO parameters for dozens of molecules concurrently in one to three weeks. In contrast, to synthesize and test the same molecules could take one to six months for each molecule.

## Electronic Paper and Molecular Cookbooks

Tomlinson's goal is to identify photoelectronically active CPs that can yield efficient solar cells. The ability to predict CP electronic properties – and to fine tune those properties by changing side groups – may allow these organic materials to perform some tricks, says Tomlinson, that doped silicon could never match.

In particular, because they are physically flexible carbon chains, CPs may offer the ability to imbed computers in everyday devices. And they can be transparent. "Smart" windows could change optical properties in response to incoming light or user commands. "Electronic paper" that includes these smart windows could combine the functions of ordinary paper with those of a computer terminal.

One avenue that the researchers are exploring arises because the side groups of semiconducting polymer chains can themselves be polymer chains. Longer chains festooning the main chain could make the polymer more soluble in organic solvents, and thus more accessible to be chemically manipulated. Long-chain side groups may also broaden the extent to which those groups could alter the properties of the entire polymer, yielding new components.

Possibly the most exciting outcome of this research could be a molecular cookbook. Tomlinson envisions a database that would allow device designers to stipulate the physical, electronic and photonic behavior they need, and from that generate a list of compounds that would meet the specifications. It's a level of programmed innovation that would revolutionize device design.

"That's a long way off," says Tomlinson, and would inevitably take the contributions of many researchers. But her work is an important step that helps the field make the crucial transition from random testing into targeted development. ■ (KC)

**More info:** www.psc.edu/science/2012/polymers/

### How Polymers Conjugate

The ability of organic molecules to act as conductors — and semiconductors — stems from the physics that pulls atoms together into molecules. When two atoms share one of their electrons with each other, they form a single bond; if two atoms share two electrons each, they form a double bond — indicated in the bonding diagram (left) by two bars.

When a chain of carbons alternates between single and double bonds, as in the bonding diagram, the system becomes "conjugated" — bonded in a manner in which an electron from each carbon atom gains the ability to "come loose." These electrons then are no longer tied to the atoms they came from — they can move freely along the chain of carbons, and the polymer then can act as a molecule-sized wire.

# BRIGHT LIGHTS, BIG COSMOS

## Relying on Blacklight's shared memory, a team of astrophysicists is running the most sophisticated, largest simulations yet undertaken of when the cosmos first began to blaze with islands of light

Before there was a Milky Way galaxy, a solar system or planet Earth, the Universe – as if taking a nap after the birth effort of the Big Bang – was wrapped in a blanket of cosmic fog. There were as yet no stars nor galaxies. Cosmologically speaking, it was the Dark Ages.

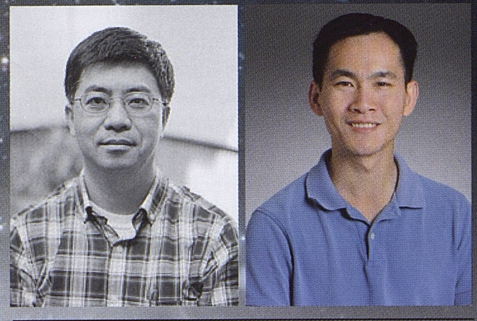
Initially, in that first mysterious microsecond about 13.7 billion years ago, there was light. And then an instant after the Big Bang, as the prevailing cosmological theory is often called, matter was an expanding soup of elementary particles, quarks and gluons and photons, which in turn evolved into a plasma, an ultra-hot swirl of protons, neutrons and electrons – with temperatures too hot for atoms to form. As the plasma cooled and the rapidly growing baby Universe was still very young – around 380,000 years, protons and electrons came together and made neutral hydrogen atoms.

The cosmos had recombined, and by this time the unimaginably hot initial spark had ballooned into immensity and cooled to about 3,000° Kelvin. Besides the dimming echoes of the Big Bang, the cosmic background radiation, nothing was hot enough to radiate. The Dark Ages passed, obviously. But how did light come back into the cosmos?

"This is one of major frontiers in astrophysical research," says Princeton University theoretical physicist and cosmologist Renyue Cen. "When did the first stars, black holes, galaxies and quasars form? These questions are fundamentally important."

Cen and Carnegie Mellon cosmologist Hy Trac lead a team of physicists – including post-doctoral researchers Nick Battaglia and Aravind Natarajan and graduate student Paul La Plante – undertaking

Artist's impression of a primordial quasar as it might have been, surrounded by sheets of gas, dust, stars and early star clusters. Image credit: NASA/ESA/ESO/Wolfram Freudling et al. (STECF)



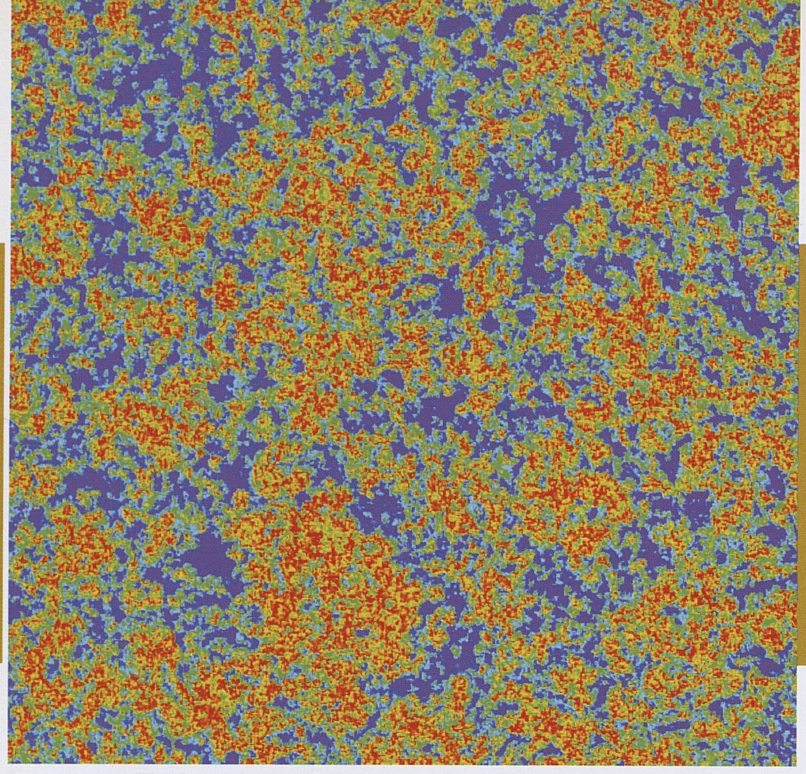
▲ **Renyue Cen,** Princeton University (left) and **Hy Trac,** Carnegie Mellon University

a series of very large-scale computational simulations to help answer these questions. What astrophysicists understand is that, gradually, tiny ripples in matter started a process by which gas coalesced and ultimately collapsed, under the action of gravity, to form the first stars, lit up with nuclear fusion, and quasars, powered by black holes.

"This marks the emergence of the first luminous bodies in the Universe," says Trac. Over the next few hundred million years, ultraviolet light from these first stars and galaxies converted the gas surrounding them into a much hotter, thinner plasma of protons and electrons – "reionized" it – and the Universe came to look much like it does today: a sea of blackness dotted with islands of light.

In between recombination and the start of reionization, a period that, relatively speaking, didn't last long – a few hundred million years – things happened by which the Universe began to structure itself into a vast web – sheets, filaments and knots of matter. Within this cosmic web, galaxies formed and processes originated that, over billions of years, led to living organisms and consciousness – which allows scientists in 2012 to try to understand how, out of inanimate energy and matter, we got to where we are.

"We have basically no information," says Cen. "That period of darkness to the end of reionization is a black box." To delve into this black box more deeply than has been done until now, Cen and Trac turned to a sophisticated software approach, RADHYDRO, which they developed for these studies. RADHYDRO incorporates physics of three mechanisms involved in shaping the cosmos as it emerged from darkness: gravity, hydrodynamics, and radiation.



**Radiation from Neutral Hydrogen**
This graphic from the simulations shows 21 centimeter radiation (increasing from blue to red, in thousandths of degrees Kelvin) emitted by neutral hydrogen in a 5° x 5° map of the sky at a time when the Universe was approximately 500 million years old. Upcoming radio experiments such as the Low Frequency Array and the Square Kilometer Array will measure these neutral hydrogen regions.

Using PSC's Blacklight, specifically because of its large amount of shared memory, the researchers have simulated a larger chunk of the reionizing Universe than previously attempted. Their simulations, still underway, have begun to zero-in on spectral signatures – the imprint of electrons released by reionization on the cosmic microwave background radiation and a signal (the "21 centimeter line") produced by neutral hydrogen atoms.

More precise information on these signatures of reionization will help to guide several large-scale observation efforts soon to be up and running. Some are space-based – the Planck Space Observatory, launched by the European Space Agency, and NASA's James Webb Space Telescope – and others are precision ground-based telescopes. "We are making maps of the sky," says Trac, "at various wavelengths and calculating theoretical predictions to compare with observational data."

## A Three-in-One Model

Even with the most powerful supercomputers, it isn't possible to model every atom, proton and photon of light in the entire Universe. The researchers, nevertheless, have taken their work beyond previous efforts at modeling cosmic reionization. Their innovative software RADHYDRO is more comprehensive in the physics it incorporates than prior models, and they are modeling a larger volume of space with higher resolution in the quantity of particles and light rays they represent within that volume.

RADHYDRO includes gravity, which takes into account the invisible substance that comprises most of the mass in the Universe. "Eighty-five percent of the Universe is in the form of dark matter," says Trac, which interacts with other matter – including the protons, neutrons, and electrons that make up the visible Universe – only through gravity.

***"Blacklight makes it possible for us to run the largest simulations of reionization in the world."***

With hydrodynamics, RADHYDRO takes account of cosmic gases, primarily hydrogen and helium, by tracking their evolution as a fluid. At this very large scale, says Trac, rather than thinking of gas as individual hydrogen and helium atoms, it can be effectively treated as an ideal fluid. The microscopic interactions are accounted for in the fluid equations that describe macroscopic properties, including extremes in pressure experienced by gases in space.

RADHYDRO's third component, radiation, distinguishes this code from most other cosmological modeling, which generally doesn't include the physics of electromagnetic radiation. As emitted from normal matter in space, radiation influences the evolution of the reionizing Universe. "We use radiative-transfer algorithms," says Trac, "that follow the propagation of radiation from early stars and galaxies out into expanding spacetime."

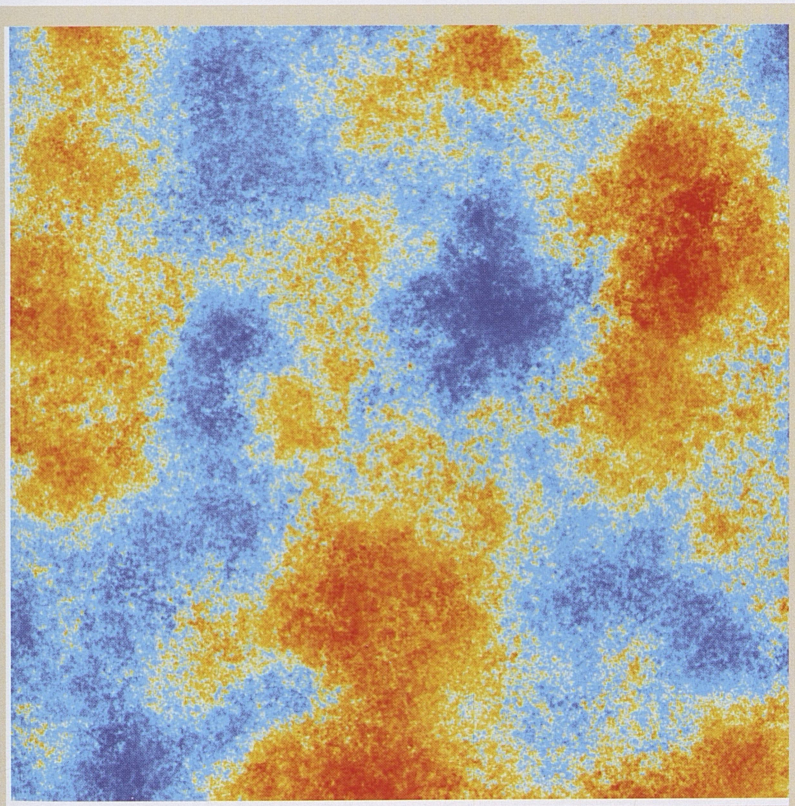
In volume, the calculations are almost unimaginably vast. Their most recent simulations cover a cube of 143 million "parsecs" – a little under 500 million light years – on each side. In miles, that's three followed by 21 zeros. This compares with a diameter of the Universe at the time of 12 billion parsecs. The simulation, then, is 1.25 percent the width of the reionizing cosmos. About right, says Cen, to capture the large-scale graininess of the Universe at that point.

Their simulations aren't just the largest yet attempted for this period of the cosmos. They are also highly detailed. Their virtual box contains eight billion particles of dark matter, eight billion gas elements, and two billion light rays. "The more particles you have," notes Cen, "the more resolving power."

## Traces of the Dark Ages

Princeton cosmologist Jeremiah Ostriker isn't involved in this project but is a pioneer in simulating this epoch of the cosmos. The difference in complexity and detail between his own earlier work (first with Cen and then with Princeton collaborator Weihsueh A. Chiu) and Cen and Trac's simulations, he says, "is the difference between modeling traffic with bumper cars and modeling it with all the detail of a superhighway. Among cosmology models, this is a real detailed test. When we didn't have this model and tried to make predictions, we obtained much less accurate answers."

Although their work is still underway, Cen and Trac and their colleagues have produced several papers. "For the first few papers in this series," says Trac, "we are describing the method and studying how various observable phenomena change when we alter the reionization process." This will help theorists understand data coming in from current and future telescopes such as the Atacama Cosmology Telescope, Planck Space Observatory, the Low Frequency Array, and the Square Kilometer Array.



▲ **The Cosmic Microwave Background**
This graphic from the simulations shows temperature fluctuations in the cosmic microwave background (CMB) radiation generated due to cosmic photons scattering with fast-moving electrons during the epoch of reionization. This temperature signal, which registers in millionths of degrees Kelvin, can be positive (red) or negative (blue) depending on whether the electrons are moving toward or away from us. The square represents a 15° x 15° map of the sky that spans a time period from when the Universe was approximately 200 million to one billion years old. Ongoing experiments such as the Atacama Cosmology Telescope and Planck Space Observatory will measure these temperature distortions in the CMB.

A key parameter the simulations track is how photons from the cosmic microwave background (CMB) – the low-frequency, microwave rumbles of the Big Bang – scatter against free electrons. By definition, reionization liberates electrons – splitting the neutral hydrogen of the Dark Ages, making it possible to study, says Trac, "the imprints of reionization on the CMB temperature and polarization." Another key parameter is a signal neutral hydrogen emits at a wavelength of 21 centimeters. The researchers are beginning to close-in on how observations of this 21 centimeter radiation will constrain how and when reionization must have occurred.

The researchers are working to scale up their model – to run efficiently on more processors, for which PSC staff have been crucial. In particular because of RADHYDRO's incorporation of radiation physics, Blacklight, with its large shared memory, is the best possible machine for this work. PSC scientists Roberto Gomez and Rick Costa helped to overcome obstacles in efficiently using software called OpenMP, which allows the software to communicate among processors. "Because these photons are always moving," says Trac, "communicating them between different processors is very difficult. Blacklight makes it possible for us to run the largest simulations of reionization in the world."

Cen and Trac have run RADHYDRO on Blacklight efficiently with as many as 512 processors and are working to use 2,048. With scaling up the simulation, the researchers plan to include three times more particles and a larger volume; the 2.048 processor simulation will include 29 billion dark matter particles, 29 billion gas elements, and 17 billion light rays. "We'd like to have the ability to resolve small dark matter halos where small galaxies form and reside," says Cen, "which were the bulk of the luminous galaxies at this period."

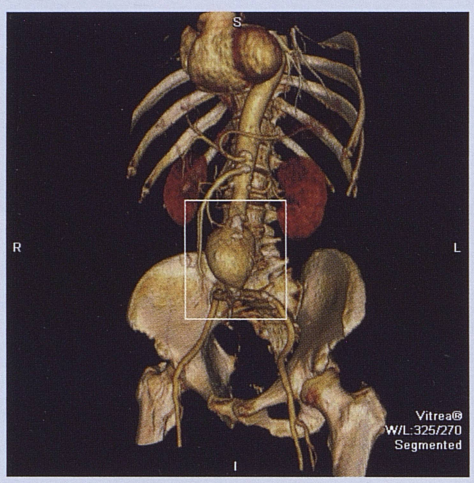
With increased resolution, the simulations will, for instance, help to determine what kind of luminous matter first began radiating – stars or black holes – as the Universe reionized. Stars and black holes leave different imprints, says Cen, and the simulations can provide concrete clues that will help to optimize the search strategies of space and ground-based observation.

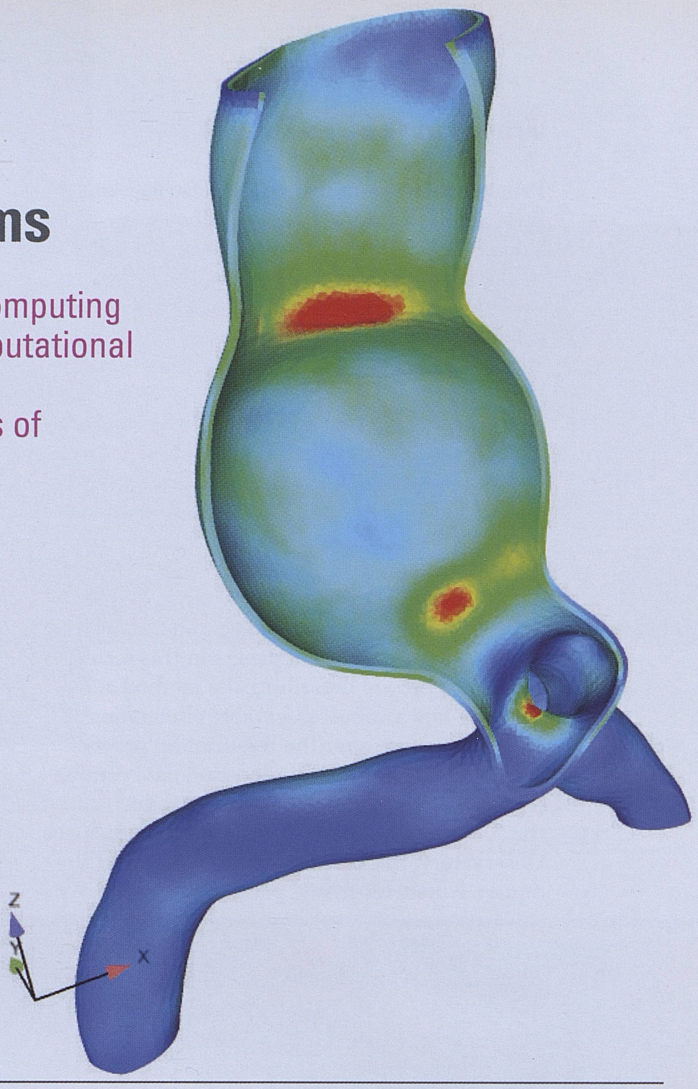
Eventually, telescope observations and simulations will feed insights back and forth, helping the other to advance. By knowing what we're looking at, we can better understand how the early Universe worked – and how it became the collection of bright islands of shining stars and galaxies we see today. ■ (KC)

**More info:** www.psc.edu/science/2012/cosmos/

# Modeling Aortic Aneurysms

## With help from XSEDE consulting and computing resources, researchers have done computational modeling of the biomechanics of aortic aneurysms starting from medical images of individual patients




The cut-section graphic (right), from modeling by Finol and colleagues, zooms into the interior of an abdominal aortic aneurysm, showing wall thickness and indicating wall stress (increasing from blue to red). "We have software to make computational models from medical images of individual patients," says Finol, "which takes into account their aortic wall thickness, slice by slice, *in vivo*, and from that to predict wall-stress distribution. No one else has done this before with this level of accuracy."

Abdominal aortic aneurysm (AAA), an enlargement of the abdominal aorta by 50 percent or more, occurs in more than 8 percent of people over 65. It can lead to fatal rupture and is the tenth-leading cause of death for men over 50. Current medical practice lacks the ability to fully assess AAA risk of rupture, with one of the known factors being AAA wall stress, for which there are no reliable *in vivo* measurement techniques.

Many key parameters of AAAs, furthermore, show wide variation among individuals. Ender Finol, director of the Vascular Biomechanics and Biofluids Laboratory at The University of Texas at San Antonio, has developed computational protocols (finite-element analysis using ADINA software) for modeling patient-specific AAA features so that they can be translated to reliable individualized wall-stress predictions. The goal is to use this modeling for to assess of the risk for rupture for individual patients, and thereby to help guide decisions about surgical intervention.

Finol, who until a year ago was at Carnegie Mellon, collaborates with Pittsburgh's Allegheny General Hospital in gathering imaging data of AAA patients. XSEDE consultant Anirban Jana of PSC has provided advice on ADINA options and coding (with software called MATLAB) and in developing a method to initialize the model – to set its boundary conditions – from patient-specific profiles. "Anirban's role has been more than just to help launch our finite-element models," says Finol. "He has been a member of the advisory committee for my doctoral student, Samarth Raut, who did much of our AAA modeling. Anirban has provided valuable input regularly to this project."

With computations on XSEDE's Pople (now decommissioned) and Blacklight, Finol has presented conference papers (with Jana as co-author) on computational solid-stress (CSS) modeling of patient-specific AAAs. Results show wall stresses more sensitive to changes in AAA shape, and the work, further, suggests that rupture risk may be characterized in relation to AAA morphology. In ongoing work with Blacklight, Finol is increasing the number of patient-specific AAA cases modeled, with the aim of completing analysis from 200 individual AAAs, each of which requires geometry reconstruction and meshing with nearly three-million degrees of freedom for a CSS simulation. Using the shared-memory version of ADINA, Jana has found that the problem optimizes at eight cores with up to 32 cores for faster time to solution.

# When Small Worlds Collide

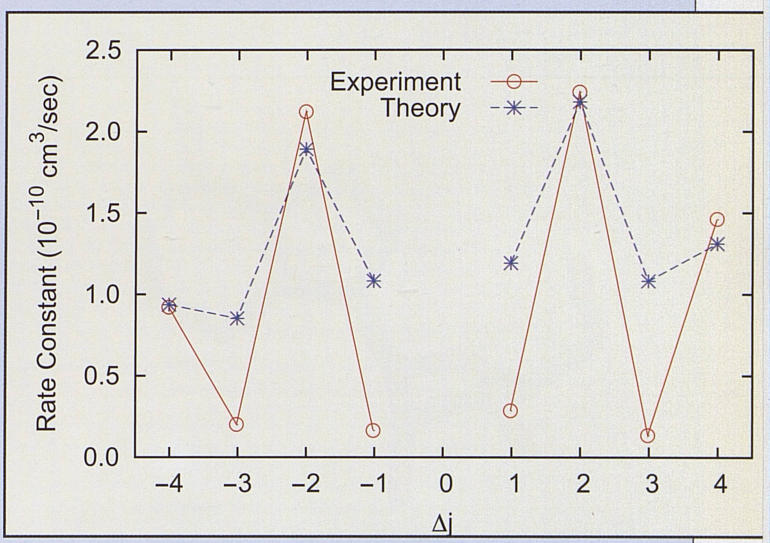
## With the advantage of large memory per processor on Blacklight, researchers are learning about how collisions alter spin in the quantum world

Imagine that you're a fastball pitcher, but instead of a baseball you're hurling a helium atom toward, not a batter's swinging bat, but a spinning molecule of sodium-potassium, NaK. You wouldn't be able to see what happens, but the impact releases energy from NaK and changes its orientation in space. "Some collisions are head-on," says physicist A. Peet Hickman of Lehigh University, "and some are glancing blows."

Hickman and colleagues are using Blacklight to do calculations that model this kind of collision, to see how the impact changes quantum properties of the rotating NaK. Their calculations go hand-in-hand with experiments of John Huennekens and his group in Lehigh's Department of Physics, using a polarized laser beam to rotate NaK molecules as they are smashed into by helium atoms (or in other experiments, atoms of argon or potassium). The Lehigh experiments measure rates of collision that change NaK's rotational quantum number J.

It's pure research, gaining fundamental knowledge about the topsy-turviness of how things happen in the quantum world, where the rules of the macroscopic every-day world we live in don't apply. This research holds possibilities for "quantum computing" in which changing the spin of an atom, switching from one angular momentum to another, could be a way of storing binary information at much higher density than current technologies.

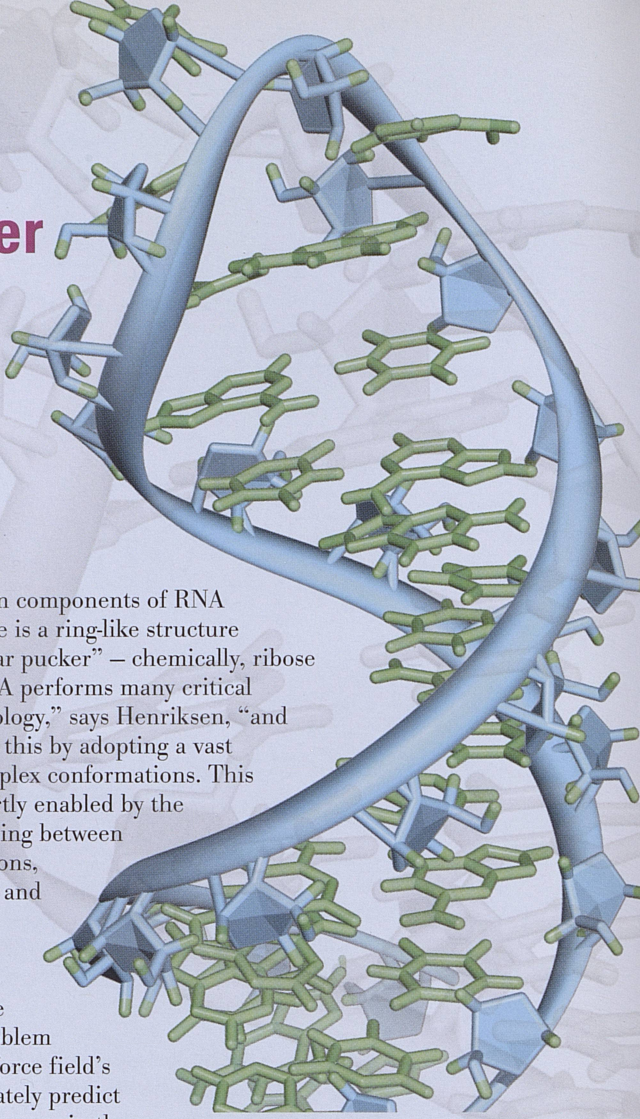
The researchers used Blacklight (with code that Hickman wrote) for extensive calculations that use the laws of quantum mechanics to determine the outcome of the collisions. Hickman also used XSEDE resources at Texas, Illinois and San Diego for electronic-structure calculations that show how the energy properties of NaK depend on the bond length between Na and K. Their theoretical results are in good agreement with some of the main features of the experimental results (see figure), and provide several predictions that can be tested in future experimental work, among them that the vibrational level has a significant effect on the rate of rotational transitions. Experimental work is underway to test this prediction. Further calculations have shown that orientation of the NaK molecules tends to be preserved in collisions with helium, even when the rotational energy changes significantly, and that this effect is very sensitive to the vibrational level. The calculations involved solving several hundred coupled differential equations, and Blacklight (which achieved 87 percent memory utilization on 96 cores) was particularly valuable, says Hickman, because of its large memory per processor.




Comparison of experimental and theoretical rate constants for rotationally inelastic scattering of helium and sodium-potassium at 600° Kelvin. These preliminary theoretical calculations were carried out for vibrational level (v) = 15. In agreement with experiment, rates for inelastic transitions with an even change in rotational quantum number (ΔJ) were found to be larger than those with odd ΔJ.

# Force Field of the Sugar Pucker

## In work aimed at developing drugs to knockout viral diseases, researchers are using Blacklight and other XSEDE resources for precise quantum calculations of the "force field" of RNA



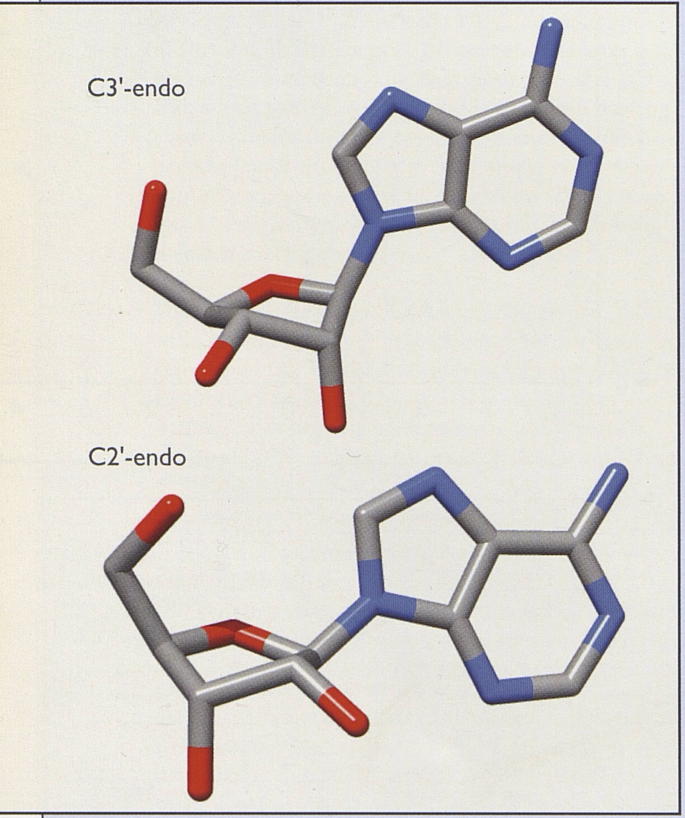
To find drugs that can deliver a knockout punch to a virus is a Mount Everest research problem. The intense effort spurred by the AIDS epidemic brought into being a small arsenal of anti-viral drugs – effective at managing symptoms, but far from a cure. For many viruses with fatal implications for humans, to go for the jugular means going for the RNA, the molecule by which most viruses – including HIV, hepatitis C, yellow fever and others – insert their genetic material and take over the cell's replication machinery.

University of Utah computational biochemist Tom Cheatham leads a team of researchers who use computational modeling to help design new therapeutic drugs. Their progress depends on the accuracy of "force fields" to model the structure of biomolecules. A recent focus of Cheatham and graduate-student colleague Niel Henriksen is the "sugar pucker" of RNA.

One of the main components of RNA helical structure is a ring-like structure called the "sugar pucker" – chemically, ribose ($C_{10}H_5O_5$). "RNA performs many critical functions in biology," says Henriksen, "and it accomplishes this by adopting a vast number of complex conformations. This flexibility is partly enabled by the sugar ring flipping between two conformations, called C3'-endo and C2'-endo."

Cheatham and Henriksen have identified a problem with the RNA force field's ability to accurately predict the energy difference in the sugar pucker flipping between the two differently oriented structures. They have used XSEDE resources, both Kraken at NICS in Tennessee and PSC's Blacklight, to address this problem. For some calculations, says Henriksen, Kraken offers the advantage of many thousands of processors. Blacklight has proven to be especially useful for quantum-mechanical calculations, which require huge amounts of memory. "The different XSEDE machines," says Henriksen, "have different strong points, but Blacklight is suited nicely for this work."

With Blacklight, Cheatham and Henriksen calculated highly accurate values for the C3'-endo to C2'-endo transition. They are in process of refining the force fields, and they are validating their findings to demonstrate that they represent substantial improvement. "We hope these improvements will allow more accurate investigations of RNA," says Henriksen, "and lead to new drug therapies targeting RNA."



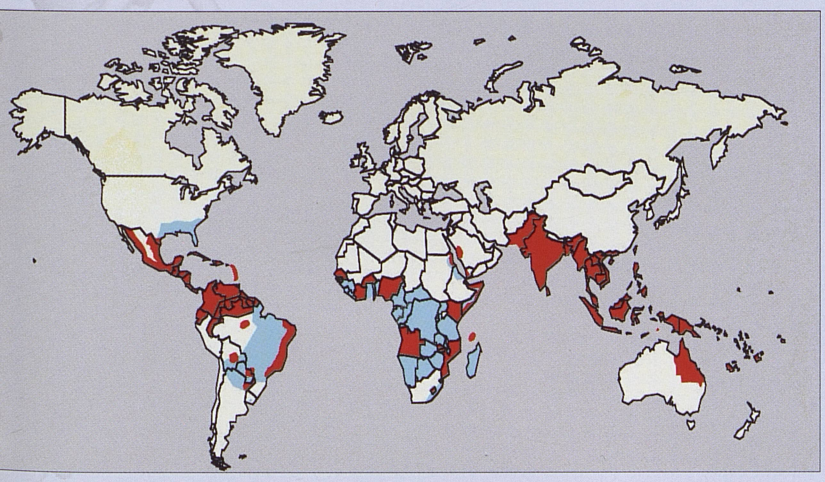

**The Sugar Pucker of RNA**
The transition that occurs between these two conformations of the "sugar pucker" in RNA, C3'-endo (top) and C2'-endo, are critical in accurate structural models of RNA, which can lead to new drug therapies to defeat viral diseases such as AIDS, hepatitis C and yellow fever.

# Fighting Dengue Resurgence

## PSC researchers are collaborating with the University of Pittsburgh MIDAS National Center of Excellence to develop epidemiological modeling that can help to control the spread of dengue fever

Dengue fever isn't near the top of the mind, generally speaking, when people think about major health problems. This tropical disease, transmitted by mosquito (*Aedes aegypti*), seemed in the 1970s to be contained in the western hemisphere and well on the way to eradication, thanks mainly to programs that reduced the standing-water breeding habitat of the mosquitoes. With world travel and the growth of urban populations, however, dengue has come back with a vengeance, including rapidly growing incidence in Africa, Asia and Brazil.

PSC scientists Nathan Stone and Shawn Brown are collaborating with the National Institutes of Health MIDAS (Models of Infectious Disease Agent Study) Center of Excellence, led by Donald Burke of the University of Pittsburgh Graduate School of Public Health, to develop computational modeling that can help to determine the impact of potential interventions on the spread of dengue. In this work, Stone and Brown also collaborate with Derek Cummings of Johns Hopkins University and John Greffenstette of the University of Pittsburgh Graduate School of Public Health.



World distribution of dengue fever, 2006, showing areas infested with *Aedes aegypti* (blue) and areas with *A. aegypti* and recent epidemic dengue fever (red).

*Aedes aegypti* mosquito biting a human

The World Health Organization (WHO) now estimates that half the world's population is at risk for dengue. The mosquito-borne dengue virus, says the WHO, infects 50 to 100 million people every year, with tens of thousands of fatal cases. In some areas dengue has become epidemic, such as Thailand, where about 70 percent of the population are either carriers or symptomatic. Symptoms, along with fever, include headaches and muscle and joint pain; for a small proportion of dengue infections, most commonly the second exposure, a hemorrhagic fever can lead to bleeding and sometimes fatal low blood pressure. Currently there is no real therapy – beyond aspirin, rest and fluid replacement– and no vaccine; to stop transmission is the only effective way to prevent dengue epidemics.

Stone and his collaborators are working to develop mathematical modeling that can show how human organization – from houses and communities to nations and continents – and movement patterns can have an impact on the spread of dengue. Their model, called CLARA – for Clara Ludlow, who helped to pioneer study of mosquitoes as a disease vector – incorporates features that go beyond prior epidemiological modeling. These include complete information on the *A. aegypti* mosquito life cycle, from eggs to larvae to adult. CLARA, notes Stone, also includes genetic information on *A. aegypti*, which bears on research on the effectiveness of dengue intervention with a bacterium, called Wolbachia, that shortens the life span of *A. aegypti*. "No other model," says Stone, "includes all these features."

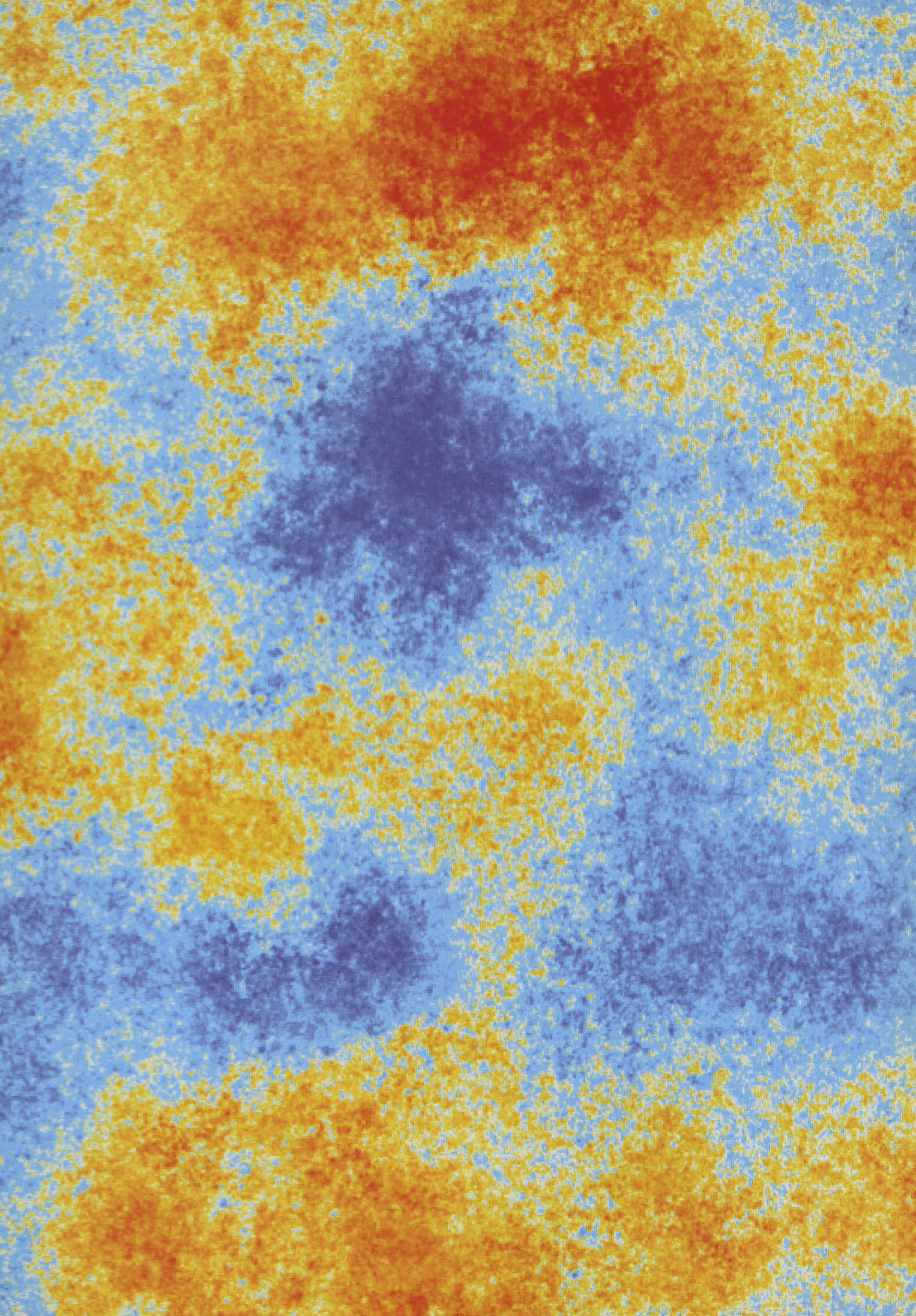
The Pittsburgh Supercomputing Center is a joint effort of Carnegie Mellon University and the University of Pittsburgh together with Westinghouse Electric Company. It was established in 1986 and is supported by several federal agencies, the Commonwealth of Pennsylvania and private industry.

PSC gratefully acknowledges significant support from the following:

The Commonwealth of Pennsylvania
The National Science Foundation
The National Institutes of Health
The National Energy Technology Laboratory
The National Oceanographic and Atmospheric Administration
The National Archives and Records Administration
The U. S. Department of Defense
The U. S. Department of Energy
D. E. Shaw Research
Cisco Systems, Inc.
Cray Inc.
DataDirect Networks
DSF Charitable Foundation
The Grable Foundation
Microsoft Corporation
Silicon Graphics, Inc.
The Buhl Foundation
Bill and Melinda Gates Foundation

EDITOR/WRITER: Michael Schneider, PSC
CONTRIBUTING WRITER: Ken Chiacchia
DESIGN: Shandra Williams, PSC
PRODUCTION COORDINATOR: Vivian Benton, PSC
TRANSCRIBING: PSC hotline staff
PHOTOGRAPHY: Tim Kaulen & Jordan Bush, Photography & Graphic Services at Mellon Institute.
GRAPHICS: Thanks to the researchers & PSC scientists.
COVER GRAPHIC: Image of an abdominal aortic aneurysm, from modeling by Ender Finol, with color showing wall stress, increasing from blue to red, see p. 44.
PRINTING: Hoechstetter Printing

Printed on Sappy McCoy Paper, a premium sheet with 10 percent post-consumer waste fiber, with vegetable-based inks.



PITTSBURGH SUPERCOMPUTING CENTER
300 S. CRAIG STREET
PITTSBURGH, PENNSYLVANIA 15213