XSEDE

The Extreme Science and Engineering Discovery Environment

2011-2012 Annual Highlights

XSEDE

The Extreme Science and Engineering Discovery Environment 2011-2012 Annual Highlights

The XSEDE project is lowering barriers and breaking them. At the end of the first year of the project, XSEDE has made a marked impact on science and engineering research, training, and education and outreach efforts. Researchers in a wide variety of fields who are looking for various computing capabilities are finding that XSEDE facilitates greater access to these advanced digital services and provides expertise to most effectively harness them.

With the help of XSEDE, thousands of researchers are improving disease prevention and diagnosis, ensuring effective drug delivery, establishing reliable energy sources, providing greater access to historical information, facilitating communities of scientific study in genomics and other life sciences, developing a broader base of knowledge on galaxy and star formation, understanding global climate change and its connection with water issues, impacting the world of finance, and encouraging more students and minorities to play a part in advancements in science, technology, engineering, and mathematics.

XSEDE is establishing the course of future research and cyberinfrastructure development by bringing more researchers into the realm of advanced computing and data-driven research, enhancing productivity, and facilitating a wider array of discovery and knowledge generation.

XSEDE is a five-year, $121 million project supported by the National Science Foundation.

John Towns
XSEDE Project Director

Welcome to the annual highlights publication of the Extreme Science and Engineering Discovery Environment. As the first issue that is firmly grounded in the work of XSEDE, it is exciting and gratifying to see the stories that have been selected. There was a tremendous portfolio of articles from which to choose. This issue represents the work of more than 9,000 researchers participating in more than 2,000 projects, producing more than 2,000 publications during the past year.

XSEDE's first year included a tremendous amount of behind-the-scenes work to establish the project with vastly improved internal processes and transparency to the research communities it supports. We developed the XSEDE User Portal as the single interface to our environment, further developed our connections to campuses, and initiated new workforce development efforts. The range of activities has been incredible. It has set the stage for us to begin to deliver a suite of new capabilities and services during our second year and improve significantly on the reliability and availability of the services we continue to support. We have begun to roll out a series of documents on our website that provide detailed technical information on XSEDE plans and designs, and we established many new engineering and quality assurance processes.

A year ago we were excited about the start of a new project; now we have an even greater excitement, as we have completed most of the groundwork necessary to facilitate a broad range of ground-breaking research. New or improved software products will be delivered to the community on a regular basis, ranging from a wide area filesystem deployment to a new user allocations request interface to a new ticket system. These will be enhanced by training on new resources with new technologies and formal adoption of undergraduate and graduate certificate and degree programs.

XSEDE is transitioning from a "startup" mode to the regular delivery of value to the community, and this is truly exciting!

XSEDE

Extreme Science and Engineering Discovery Environment

XSEDE, the Extreme Science and Engineering Discovery Environment, is the most advanced, powerful and robust collection of integrated digital resources and services in the world. With singular interfaces for allocations, support, and other key services, XSEDE is a virtual system that scientists and researchers can use to interactively share computing resources, data, and expertise. XSEDE integrates the resources and services, makes them easier to use, and helps more people use them.

XSEDE is led by the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign.

The partnership includes:

- Cornell University Center for Advanced Computing
- Indiana University
- Jülich Supercomputing Centre
- National Center for Atmospheric Research
- National Institute for Computational Sciences – University of Tennessee Knoxville
- Ohio Supercomputer Center - The Ohio State University
- Pittsburgh Supercomputing Center - Carnegie Mellon University/University of Pittsburgh
- Purdue University
- Rice University
- San Diego Supercomputer Center - University of California, San Diego
- Shodor Education Foundation
- Southeastern Universities Research Association
- Texas Advanced Computing Center - The University of Texas at Austin
- University of California, Berkeley
- University of Chicago
- University of Virginia

On the web: **xsede.org**

XSEDE is supported by the National Science Foundation.

On the cover:

This volume rendering (where red and blue indicates hot and cold) depicts a region 110,000 light years across — approximately the size of the disk of the Milky Way — 800 million years after the Big Bang. The largest galaxy in the simulation "(where red and blue indicates hot and cold) is only one-thousandth the mass of the Milky Way, yet it provides the majority of radiation that heats and ionizes its surroundings.

Courtesy of John H. Wise, Georgia Institute of Technology

Contents

Science Gateways Growing in Popularity among XSEDE Users

A simplified version of the "tree of life," intended to show that life on Earth shares a common, genetic history with complex origins.

Courtesy of Nick Kurzenko, Greg Rouse, and the U. S. Fish and Wildlife Service.

XSEDE13 theme, "Gateway to Discovery," highlights gateway successes

Any way one measures it — by the number of new users, the amount of compute hours requested, or the number of published research papers — the use of science gateways has surged in popularity among XSEDE's research community.

Gateways are valuable web-based tools that allow scientists, students, and others to conduct a wide range of studies in significantly shorter times and without having to understand all the complexities of larger high-performance computing (HPC) systems. Gateways also can be readily used for teaching classes, workshops, and tutorials without having to set up complex codes on HPC resources or create new accounts for participants.

The theme for XSEDE13, XSEDE's annual conference, is "Gateway to Discovery," underscoring the tremendous impact science gateways have had on the broader research community. "The ability (of gateways) to deliver the functionality of high-performance resources without the complexity has been a real win for science," says Nancy Wilkins-Diehr, XSEDE13 general chair and director of the Science Gateways program since its inception.

XSEDE's growing collection of gateways currently numbers more than 30. Among the newer gateways is CyberGIS, led by the National Center for Supercomputing Applications (NCSA), which can handle very large geospatial data sets and complex analyses that conventional geographic information systems (GIS) software do not provide.

AMP

The Asteroseismic Modeling Portal (AMP), another XSEDE gateway, provides a web-based interface for astronomers to run and view simulations that derive the properties of Sun-like stars from observations of their pulsation frequencies. AMP was developed by the High Altitude Observatory and the Computational Systems and Information Laboratory at the National Center for Atmospheric Research (NCAR). AMP serves an international community of 95 researchers organized by Travis Metcalfe, and since December 2008 it has run more than 14,000 jobs on the Kraken supercomputer at the National Institute for Computational Sciences (NICS) and more than 12,000 jobs on the supercomputers at NCAR.

Use of the AMP portal increased 138 percent in 2012, and users have reported being very satisfied with the workflow of the portal, as it simplifies running a computational experiment to several clicks, provides notification after the jobs are done, and automatically provides post-processing output data and graphics.

CIPRES Gateway

CIPRES stands for Cyber Infrastructure for Phylogenetic RESearch. It is a public resource for the community that studies evolutionary relationships among virtually every species on the planet. Created by researchers at the San Diego Supercomputer Center (SDSC) at UC San Diego, CIPRES is designed to provide researchers with access to XSEDE's computational resources through a simple browser interface.

The number of users submitting jobs to CIPRES on a monthly basis increased 400 percent from 2009 to mid-2012. Each month sees an average of 140 new users, and the number of repeat users also has increased steadily. A recent survey showed that use of the CIPRES Gateway enabled more than 400 publications, illustrating the significant impact that gateway projects can have on scientific progress.

"The CIPRES Gateway has been successful in enabling access to XSEDE HPC and cyberinfrastructure resources for users who would not otherwise be able to use them," according to Mark Miller, an SDSC researcher and principal investigator of the CIPRES Gateway. "Many of our users tell us that they could not have done their research without this access."

Grant no.: NSF EF 03-31648, NSF OCI-1053575 (XSEDE), NSF DBI-0735191, and NIH 5R01GM073931

For more information: https://www.xsede.org/gateways-overview

Story by Jan Zverina

Taming the BEAST

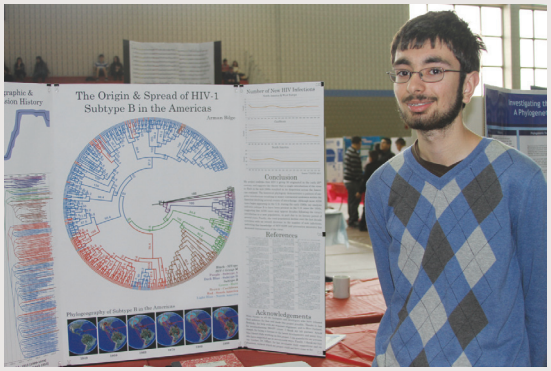
Arman Bilge, a 10th grader at Lexington High School in Massachusetts, was a newbie to phylogenetics when a science teacher organized an after-school phylogenetic tree club. In the club, Bilge learned how to use a variety of software applications, including one well known to systematic biologists called BEAST.

That led Bilge to create a map and timeline that identified when the Human Immunodeficiency Virus (HIV) arrived in the Americas and where and when it spread across North and South America.

"The BEAST is a beast," Bilge says, adding that he managed to tame it enough to create a detailed phylogenetic tree based on similarities and differences in the 3,000 nucleotide subunits of a gene for an envelope protein among 400 known HIV-1 strains.

Bilge first tried to run the analysis on his home computer. "I ran it for three weeks, but I didn't reach the accepted way of knowing that you came to the end," he says, adding that his parameters and settings were impossible for any computer to analyze. "So I started multiple simultaneous runs on CIPRES, and the geographic component of my project is the result of the concatenation of these analyses." The phylogenetic tree Bilge published for his science fair project was the one that BEAST said was the most optimal, and Bilge says his conclusions support previously published results of HIV experts in that "a single introduction of the virus in Haiti in the mid-1900s resulted in its dispersion across the American continent."

While Bilge may not have been satisfied with his residual statistical uncertainty, the judges at the 2012 Massachusetts Science and Engineering Fair were. They awarded him first place in the biology category.



Arman Bilge, Lexington High School

Superfast Gene Assembly, the Next Generation is Now

XSEDE's Extended Collaborative Support Services is providing tools that facilitate assembly and analysis of next-generation sequencing data

The colonial tuco-tuco (*Ctenomys sociabilis*) is of particular interest in behavioral genomics because behavior varies — in the same species — between social and solitary living conditions.

In genomics, the next generation is now. This relatively new branch of the life sciences has exploded in the last few years with possibility and data. "Next generation" sequencing tools have taken genomics to studies of nearly every kind of organism, by deciphering the order of nucleotide bases — A, G, C and T (adenine, guanine, cytosine and thymine) — at unprecedented speeds.

The essential difference is long versus short reads. Previous sequencers did reads of about 300 to 500 and sometimes up to 1,000 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 data as would require a year for traditional sequencers. Consequently, genomics has shifted into data-intensive overdrive, with many opportunities to do important research. While it is a blessing for the life sciences, it is a major 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, and short reads amplify the computational task — many more pieces of data must be fit together based on shorter overlaps. Imagine a jigsaw picture puzzle with 100 big pieces versus the same picture with 2,000 little pieces.

Through XSEDE's Extended Collaborative Support Services, Pittsburgh Supercomputing Center (PSC) scientist Phil Blood enhanced the genomics capabilities of XSEDE's Blacklight system, making nearly all genomics software tools available as easy-to-use pre-compiled modules. Blood's genomics work garnered attention in a journal article, *GenomeWeb* (February 1, 2012), highlighting how shared-memory systems, which can hold entire base-pair datasets in random-access memory, can facilitate assembly and analysis.

Matthew MacManes, of the University of California, Berkeley, consulted with Blood and used nearly 20 different programs in his assembly and analysis of sequence data from the colonial tuco-tuco, a burrowing rodent from Patagonia. MacManes focused on tucos because of his interest in behavioral genomics — the genetic underpinnings of social behavior. This tuco is unique in that individuals within the same species display social and anti-social extremes; some live in groups while others are solitary. MacManes identified a number of tuco genes that are expressed or not depending on social behavior.

Next-generation sequencing also has opened up studies in "metagenomics" — simultaneous analysis of genes from many organisms that co-exist in the same environment. In one such study, unimaginable only a few years ago, Blood helped researchers from Oklahoma State assemble sequencing data in soil from a sugar-cane plantation in Brazil. The goal was to find enzymes that can efficiently break down non-feed plants — such as switchgrass and wheat straw — that have the potential to yield biofuel more efficiently than feed-stock plants, such as corn. The Oklahoma State team of Mostafa Elshahed, Rolf Prade, and Brian Couger completed the largest metagenomics assembly to date, totaling 1.5 billion 100 base-pair reads. Their analysis has identified thousands of candidate enzymes, all previously unknown, that offer promise to cost-effectively degrade non-feed-stock crops to biofuel.

Monogamy and the Immune System

In another project, Matthew MacManes used the Ranger system at the Texas Advanced Computing Center to investigate genetic differences in two related species of mice, one monogamous, the other sexually promiscuous. His analysis focused on the differences in bacterial communities in the female reproductive tract. He found that the promiscuous mouse has greater bacterial diversity and, furthermore, that this greater diversity correlates with a more robust genetic expression of the immune system's ability to recognize pathogens.
A paper reporting these findings appeared in *PLoS One* (May 2012).

Researcher: Matthew MacManes, University of California, Berkeley

NIH stipend: 1F32DK093227-01

Grant no.: MCB110134

Researchers: Mostafa Elshahed, Rolf Prade, and Brian Couger, Oklahoma State University

Grant no.: MCB120049

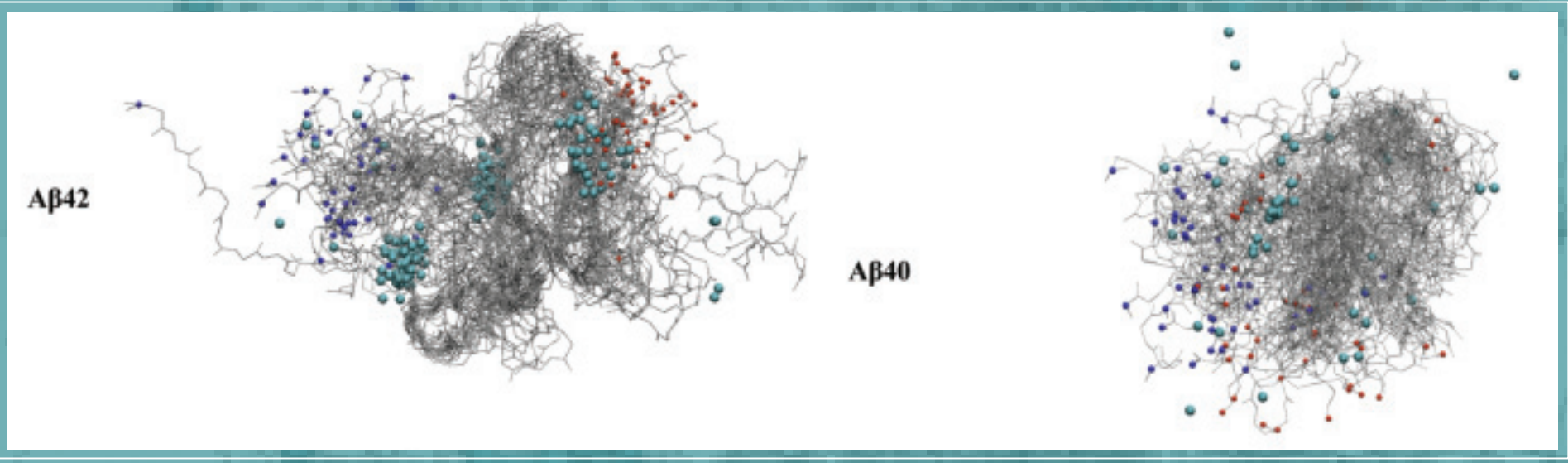
Story by Michael Schneider

Matthew MacManes, University of California, Berkeley

Brian Couger (right), Rolf Prade (center) and Tyler Werick, Oklahoma State University

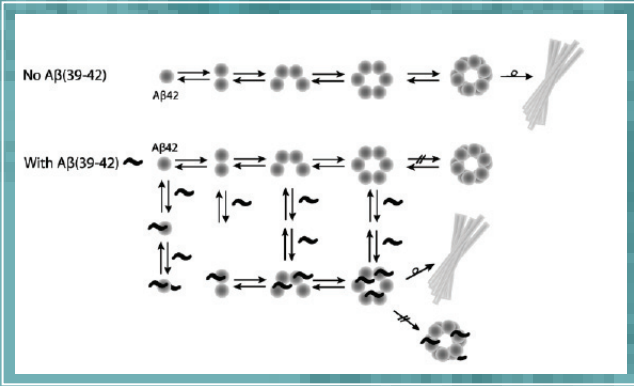
Small Molecule May Play Big Role in Alzheimer's Disease

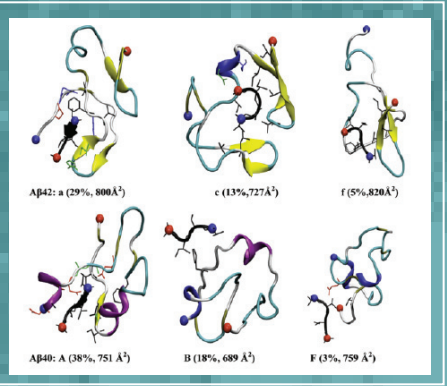
UC Santa Barbara researchers simulate amyloid fibrils on XSEDE-allocated supercomputers to improve understanding of plaque formation in the brain



These visualizations represent aspects of amyloid β-protein (Aβ) simulations that shed light on plaque formation in the brains of individuals with Alzheimer's disease.

Above: Amyloid β-protein (Aβ) binding to proteins with 40- or 42-residue peptides (Aβ40 and Aβ42, respectively).
Courtesy of The Shea Group





Normally Aβ42 forms soluble, neurotoxic oligomers before forming larger, fibrillar structures. Aβ(39-42) binds directly to Aβ42 monomer and oligomeric species eliminates the formation of large Aβ42 oligomers, driving the formation of nontoxic oligomeric species which also eventually form fibrils.
Courtesy of The Shea Group

Selected representative structures of Aβ•Aβ(39–42) complexes from the various structural families. The abundance and collision cross section are noted below each structure.
Courtesy of The Shea Group

Alzheimer's disease is one of the most dreaded and debilitating illnesses. Currently, the disease afflicts 6.5 million Americans, and the Alzheimer's Association projects it to increase to between 11 and 16 million by 2050.

"We don't know what the problem is in terms of toxicity," said Joan-Emma Shea, professor of chemistry and biochemistry at the University of California, Santa Barbara (UCSB). "This makes the disease difficult to cure."

Long knotty fibrils, formed from misfolded protein fragments, are almost always found in the brains of diseased patients. For a long time these accumulations, known as amyloid plaques, were presumed to be the cause of the disease. However, new findings support a hypothesis that the amyloid plaques are a by-product of the disease rather than the toxic agent. This paradigm shift changes the research focus to smaller, intermediate molecules that form and dissipate quickly and are difficult to isolate in brain tissue.

To study these intermediate molecules, Shea's group relies on computer models and simulations to help uncover the structure, formation, and behavior of amyloid accumulations in the brain. Since 2007, Shea has run simulations of amyloid peptides on several of XSEDE's high-performance computing systems, including Ranger and Lonestar at the Texas Advanced Computing Center and Kraken at the National Institute for Computational Sciences.

"We can identify the important structures that are adopted by these peptides at a resolution that exceeds what can be done experimentally," explains Shea. "This helps us understand which structures lead to a self-assembly."

If amyloid plaques are not the cause of the disease, what is? Shea and Michael Bowers, a professor of chemistry and biochemistry at UCSB, are taking a look at oligomers — soluble precursors of the fibrils — that could be responsible for the onset of the disease through interactions with the cell membrane.

In 2007, Shea and Bowers received a grant from the National Institutes of Health to investigate this theory. Together, they have spent the last five years looking at small peptide-based inhibitors that would prevent these oligomers from forming.

"If you can prevent the oligomers from forming, you can limit toxicity," Shea says.

In an August 2012 paper in *Biophysical Journal*, Shea and postdoctoral researcher Luca Larini studied the conformations adopted by small oligomers of peptides that are capable of aggregating into amyloid fibrils, encountered within the cell. They found that hairpin-shaped forms of the peptide initiated the aggregation of oligomers that ultimately led to the formation of a tangled fibril. Like an old slapstick routine where one person trips, another trips over them, and a pile forms, the misfolded proteins in the brain cells of those with Alzheimer's recruit other misfolded proteins and eventually grow into a large mass.

The supercomputer simulations not only have helped uncover the role of oligomers in the onset of Alzheimer's, but they are aiding in research aimed at trying to stop oligomer formation altogether. A paper in the November 2011 edition of *Biochemistry*, co-authored with postdoc Chun Wu and the Bowers group, described how a class of small molecules known as C-terminal inhibitors stopped the formation of oligomers, possibly halting disease progression before it is too late.

"Dr. Shea's simulations put a molecular face on the cross sections and oligomer distributions that we experimentally measure," said Bowers. "Of significant importance is the simulation of the Aβ42 monomer structure that very nicely correlated with our experiments."

Since 2009, the projects have required more than 13 million hours of compute time on XSEDE-allocated supercomputers. The simulations are helping researchers identify where the inhibitors bind, leading to new ideas about how inhibition can be improved.

"Nothing that we're doing here is something that we could do on our home clusters," Shea notes. "The scale of it is intractable. With growing computational resources and capabilities, we'll be able to look at how these proteins interact with membranes," adds Shea. "We're far away from simulating a whole cell, but we can start incorporating additional elements that may turn out to be important."

Joan-Emma Shea, University of California, Santa Barbara

Researcher: Joan-Emma Shea, University of California, Santa Barbara

Grant no.: 1056587

For more information:
http://web.chem.ucsb.edu/~sheagroup/
http://pubs.acs.org/doi/pdf/10.1021/bi201520b
http://www.ia.ucsb.edu/pa/display.aspx?pkey=1557

Story by Aaron Dubrow

Building Bioinformatics Expertise at Minority Universities

The Minority Access to Research Careers (MARC) program at PSC is addressing a disparity identified as a priority issue in U.S. research

Right: The 2012 MARC summer interns, with MARC program organizers Hugh Nicholas (left rear) of XSEDE and PSC; and Ricardo González (to the left of Nicholas), University of Puerto Rico; along with XSEDE scientist Alex Ropelewski (to the left of González) of PSC.

Left: Participants are busy learning about bioinformatics at the MARC 2012 Summer Institute. "We leverage the expertise and social network developed in MARC to bring novel and innovative projects from underrepresentedcommunities to XSEDE," says Sergiu Sanielevici, PSC scientist and director of XSEDE's Novel and Innovative Projects.

A training program of the National Resource for Biomedical Supercomputing at the Pittsburgh Supercomputing Center (PSC) is taking a unique proactive role toward filling the gap in scientific training at minority-serving institutions (MSIs) and historically black colleges and universities (HBCUs). Since 2001, PSC's MARC (Minority Access to Research Careers) program has evolved from providing individual training in what was at first a newly emerging discipline — bioinformatics — to a focus on the development of curricula and research programs at partner universities.

"We've implemented a multi-disciplinary course in sequence-based bioinformatics at more than 10 universities," says XSEDE scientist Hugh Nicholas of PSC, who directs the MARC program. Nicholas and his XSEDE colleague Alex Ropelewski work with five partner MSIs and HBCUs — North Carolina AT&T; University of Puerto Rico, Mayaguez; Johnson C. Smith University; Tennessee State University; and Jackson State University. The focus of the MARC effort with these MSI programs is to build a concentration or minor in bioinformatics.

"The program has shifted direction," says Ricardo González, a professor at the University of Puerto Rico, School of Medicine, who along with Nicholas leads the MARC program. "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."

Along with workshops at PSC, the MARC staff travels to partner MSIs/HBCUs 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."

Many peer-reviewed papers already have resulted from the program, notes González, work that is especially important in light of a 2011 study (*Science*, August 19, 2011) finding that black scientists were a third less likely than their white counterparts to get a research project funded. "For 10 years," González says, " the PSC program has been addressing this uneven playing field that affects black researchers."
Since 2003, the program has included a two-week workshop hosted at PSC, the MARC Summer Institute, which 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 on site 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.

Partner institutions have access to XSEDE resources, especially useful in bioinformatics as "next-generation" sequencing instruments produce skyrocketing quantities of genomics data. Tools such as Penn State's Galaxy — an open, web-based platform for data-intensive bioinformatics workflows — help to reduce the steep learning curve.

"Galaxy eliminates a lot of the complexity of getting started in this field," says Nicholas. "It has become the preferred way to perform bioinformatics analyses for both research and teaching." To facilitate this work, XSEDE recently added a high-speed network link to the main Galaxy portal at Penn State.

The MARC program sprang from a collaboration with the University of Puerto Rico School of Medicine, after González took a series of PSC workshops in the late 1990s. "This program is strong," says González," because of PSC's ability to build bridges with scientists at historically black universities and MSIs."

Researchers: Hugh Nicholas, Pittsburgh Supercomputing Center; Ricardo González, University of Puerto Rico, School of Medicine

Funding: The National Institute of General Medical Sciences, National Institutes of Health

NIH Grant no.: T36 GM095335

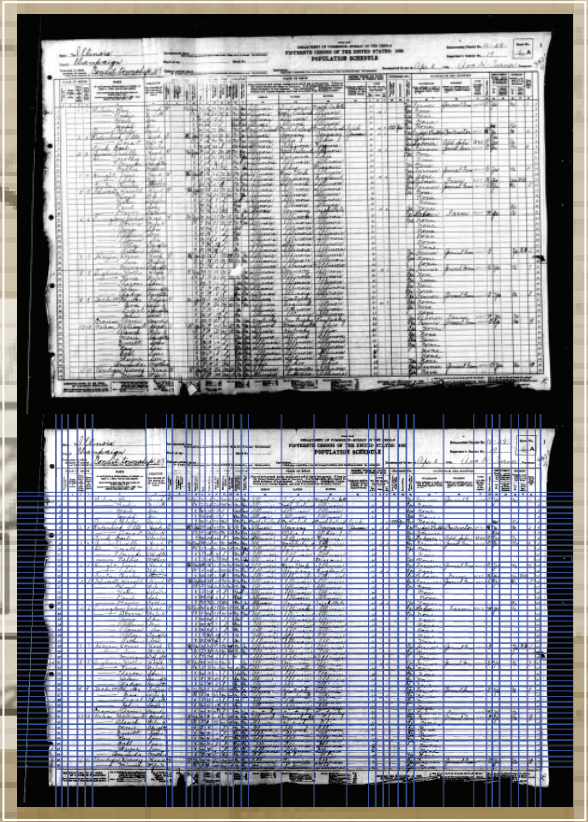
Grant no.: OCI-1053575

For more information: http://marc.psc.edu

Story by Michael Schneider

Unlocking Secrets of the Census

XSEDE resources and support are helping an NCSA team provide searchable access to a wealth of U.S. census data



Top: A handwritten Census form. There are massive amounts of information available in each handwritten form, and the NCSA team is taking up the challenge of digitizing and making the data searchable.

Bottom: By fitting a template over the Census forms, each cell can be separated into an individual image.

U.S. Census forms remain confidential for 72 years before being released to the public — a treasure trove for both genealogy buffs and researchers. The standard practice has been for the Census Bureau to create microfilm images of the millions of paper forms. Companies such as Ancestry.com then hire thousands of people to spend months transcribing the microfilm so the data can be searched and sorted online. But, starting with the release of the 1940 census data in April 2012, the Census Bureau is releasing only digital data — no more microfilm.

Of course, the millions of images constituting terabytes of data can't be easily searched for names, locations, or trends. Manual transcription of the forms is far too expensive, and optical character recognition (OCR) is not accurate enough. To maximize the usability of the digital data, the National Archives and Records Administration has provided a grant to a team based at the National Center for Supercomputing Applications (NCSA) that is developing a framework to provide "searchable access" to archives of digitized documents.

The framework enables a user to input a handwritten query — either using a mouse (or touchscreen on a mobile device) or by typing a word that then will be rendered in a handwriting font — and search a database of images of handwritten text for potential matches.

But first, significant computing power is required to pre-process the data. Each one of the millions of spreadsheet-like forms must be split into individual data cells by fitting a template over each form image. Next, each extracted cell must be converted into a numerical signature vector that roughly represents the handwritten contents of that image. Finally, a computer vision technique known as word spotting is used to compare the signature vector of the search query (such as a name, like Smith) to the signature vectors of the many, many cells, looking for similarities.

To search all 70 billion cell images would be excessively time-consuming and computationally expensive, so similar signature vectors are grouped, creating a data hierarchy that narrows the search space and returns results with reasonable speed.

Because of the vagaries of handwritten text, not all of the returned results will be perfect matches. The system's users will actually help improve the results through a passive form of crowdsourcing. After searching for "Smith," a user isn't likely to click on results that are not "Smith." The query text entered by the user can be connected to the image results the user selects, allowing the image database to be slowly annotated. Over time, validated matches can be returned to users rather than relying solely on word spotting.

An XSEDE startup allocation enabled the team to use NCSA's recently retired Ember system to develop a proof-of-concept version of the framework. An XSEDE Extended Collaborative Support Services team led by NCSA's Jay Alameda helped the group get optimal performance out of their code, assisting with mapping processes to hardware and with I/O issues. The next step is to use their allocation on the Pittsburgh Supercomputing Center's Blacklight system to index all of the 1940 census records.

Ember and Blacklight were selected because their shared-memory architecture was the best fit for the Census team's I/O-bound work.

"We deal with lots of relatively small files that need to be opened, and then we run processes that require lots of memory. Ember and Blacklight are well suited for this," McHenry says.

The work has been presented at the 2012 National Association of Government Archives and Records Administrations E-Records Forum, the 38th Annual Conference of the International Association for Social Science, the 76th Annual Meeting of the Society of American Archivists and Information Technology, and the 2012 IEEE eScience Conference.

Kenton McHenry, NCSA/University of Illinois at Urbana-Champaign

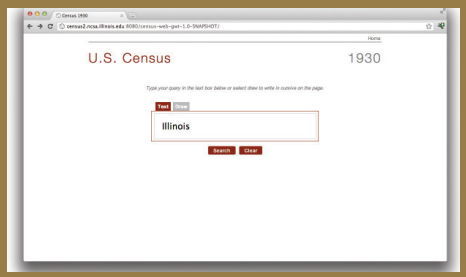
Researcher: Kenton McHenry, NCSA

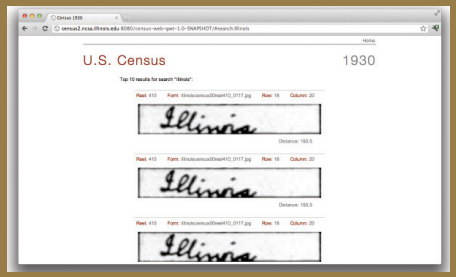
Funding: National Archives and Records Administration

Grant no.: CA SCI-9619019, 2011-2012

For more information: http://isda.ncsa.illinois.edu/drupal/project/census

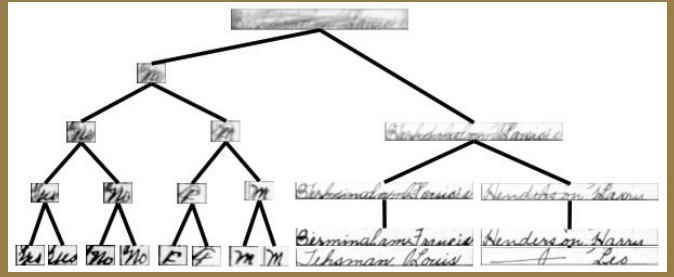
Story by Trish Barker





A snapshot of the Census framework, showing how a query can be input as typed text (which will then be rendered to look like handwriting) or handwritten. The search for "Illinois" returns an accurate match. All images: Courtesy of NCSA

Inputting a handwritten query.



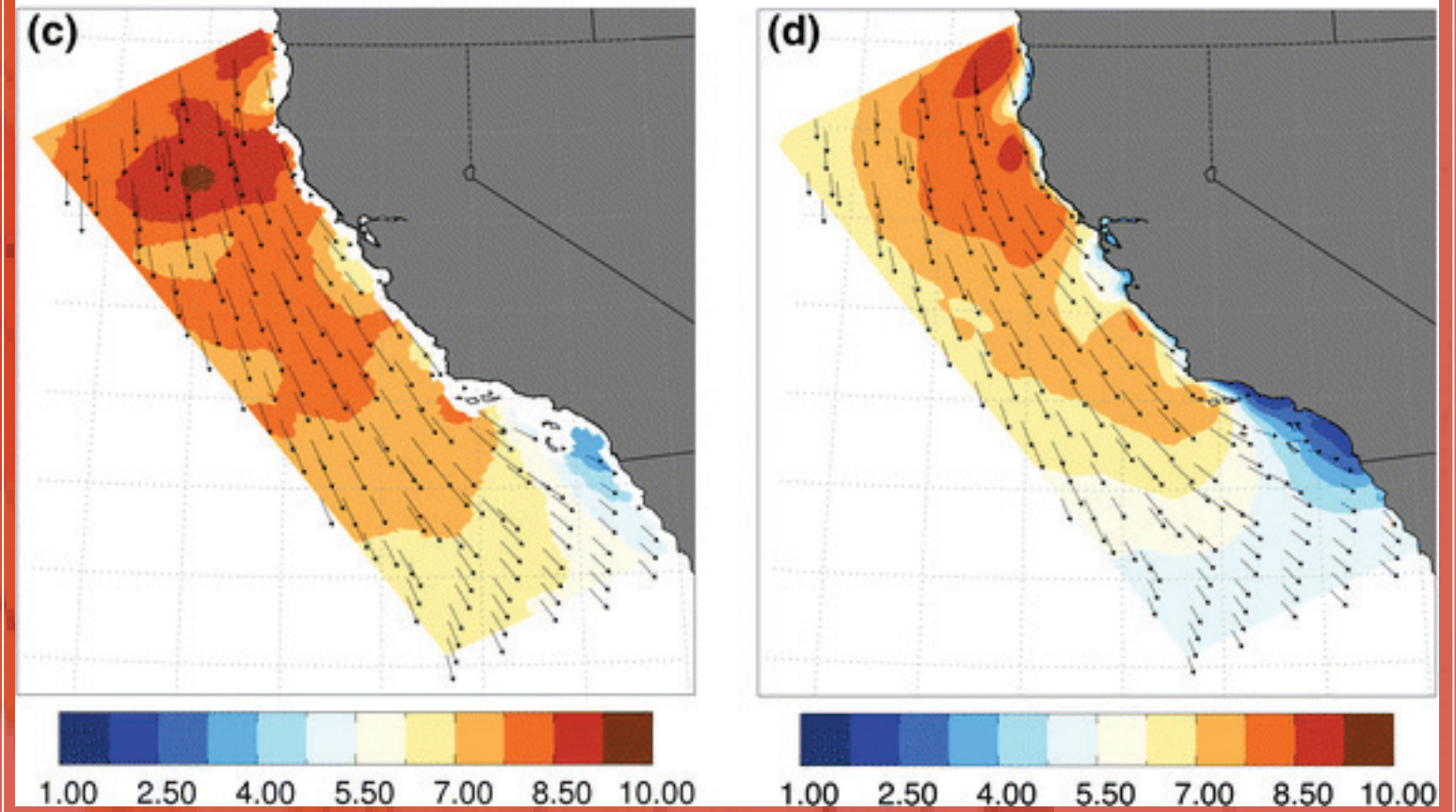
Grouping similar feature vectors narrows the search space so results can be returned more quickly.

Hot Times in Los Angeles

As part of efforts to develop a Climate Action Plan, a team of scientists produced the first study assessing the affects of climate change on a metropolitan region

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

— B. Dylan



Southern California Wind in the summer of 2002. The first graphic shows mean wind speed direction (arrows) and magnitude (decreasing from red to blue) measured by satellite, as compared to the results (left) of the Hall et al. model.

Sept. 27, 2010: The hottest day on record in Los Angeles. The official weather station thermometer broke when the temperature reached 113° F. The electrical load from the Los Angeles Department of Water and Power peaked at 6,177 megawatts, the highest in history. Was it a harbinger of LA summer days ahead?

To help get credible answers to this challenging question, Alex Hall, a professor in the University of California, Los Angeles' Institute of the Environment and Sustainability, works with a coalition of Los Angeles government, universities, and private concerns. Their aim is to develop a regional Climate Action Plan. In June 2012, Hall and his colleagues released a study, "Mid-Century Warming in the Los Angeles Region," that is the first published analysis assessing effects of global climate change at the scale of a metropolitan region.

The study predicts that, for the years 2041 to 2060, temperatures in the Los Angeles region will be higher than today's 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 foundation for the region to confront climate change. Now that we have real numbers, we can talk about adaptation."

For their modeling, the researchers relied on XSEDE's Blacklight system at the Pittsburgh Supercomputing Center and the recently retired Ember at the National Center for Supercomputing Applications, along with the National Energy Research Scientific Computing Center (in Berkeley, Calif.) and UCLA in-house computing. The project started with global models, but because this modeling — at a resolution of roughly 100 square kilometers — is too coarse to provide meaningful information at a regional scale, Hall and his colleagues did extended calculations that downscale the global models to very high resolution (about two kilometers) for the Los Angeles metropolitan area.

Among other factors, Hall's work takes into account how the coastal Pacific Ocean affects conditions over the Sierra Mountains not far inland — natural features that, among other effects, shape winds known as the Santa Anas, which have fueled some of the most furious wildfires to occur in densely populated areas. "The Santa Ana phenomenon," says Hall, "among other factors, isn't represented at all in the coarse resolution global models." Results from this modeling, reported in *Climate Dynamics* (2011), lend confidence to Hall's approach. "We can reproduce rain events when they actually occurred going back as far as data is available."

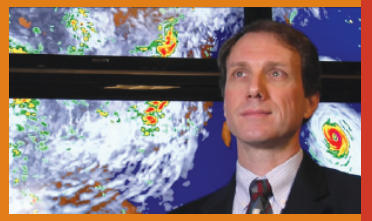
In ongoing work, Hall and his colleagues are investigating other uncertainties associated with Los Angeles regional climate-change scenarios, including critical issues — such as snowpack and low clouds — associated with water resources. "With supercomputing," says Hall, "we can simulate these phenomena in detail and see why they change and assess the credibility of these changes."

Will Drought Become the Norm?

Summer 2012: The hottest July on record for the continental United States, with repercussions still felt in food prices worldwide: Was it an anomaly of weather, or an indicator of global climate? For many, perhaps especially U.S. farmers, this was (pun intended) the burning question of the year.

In recent work, James Kinter, director of the Center for Ocean-Land-Atmosphere Studies at the Institute of Global Environment and Society, and an international group of researchers used XSEDE-allocated resources (Athena, the 166-teraflop Cray XT4 at the National Institute for Computational Sciences) to simulate global climate at the spatial resolution normally used for producing 10-day weather forecasts. Their findings, reported in *The Journal of Hydrometeorology* (August 2012), suggest trouble ahead, as Kinter described in his talk at the XSEDE12 conference in Chicago in July.

"The pattern of increasing probability of extreme drought in our simulations is quite similar to the summer 2012 drought severity map," states Kinter. "Our results suggest that, while the 2012 event itself cannot be ascribed to human-induced climate change *per se*, the severe situation we are experiencing is likely to become entirely too commonplace in the future."



Jim Kinter, Center for Ocean-Land-Atmosphere Studies.
Courtesy National Science Foundation.

Alex Hall, University of California, Los Angeles. Hall credits PSC scientist and XSEDE consultant David O'Neal, who supported the LA research team in their use of XSEDE-allocated resources. "He was extremely knowledgeable and professional. We have problems sometimes, and to have someone like him easily accessible is very helpful."

Researcher: Alex Hall, University of California, Los Angeles

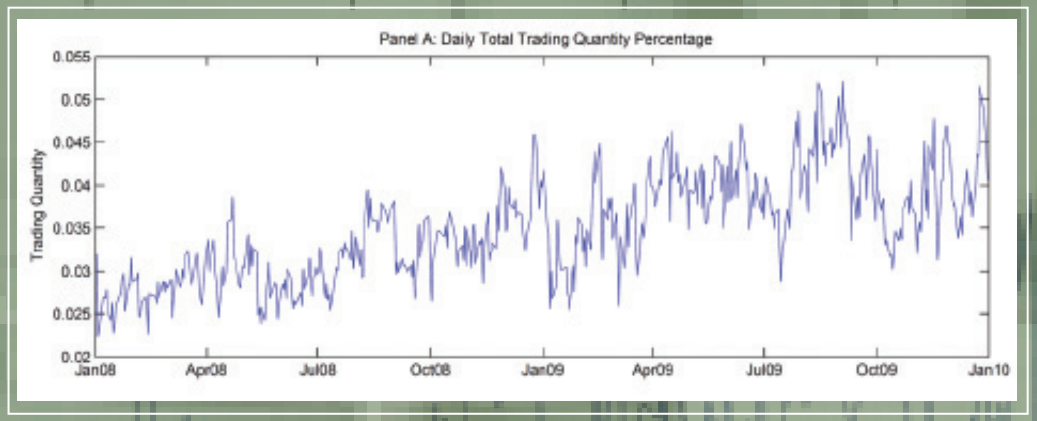
Grant no.: ATM080018N

For more information: http://c-change.la

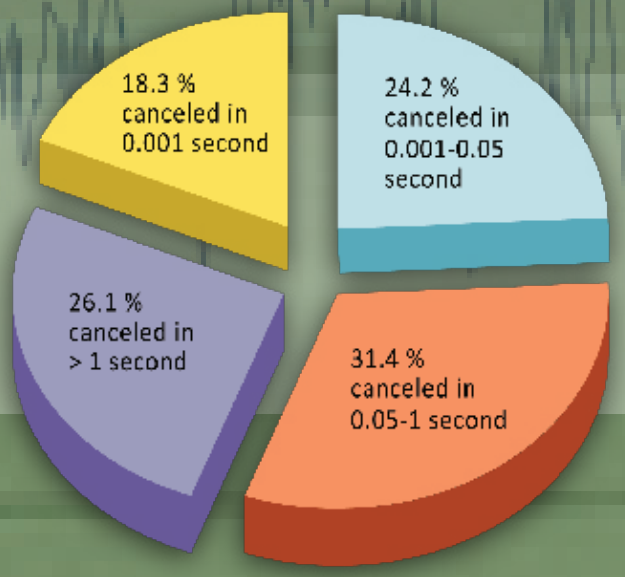
Story by Michael Schneider

Catching Up with Wall Street

Using XSEDE-allocated resources, researchers are showing how the rapid speed of computerized stock trading has little-understood, non-beneficial effects on the market



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.



On August 30, 2011, about 3 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 canceled 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.

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, 2012) 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 even 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 during 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.

HFT has happened so quickly that regulators 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.

One of the first problems researchers face 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, marking the biggest one-day drop in its history. Analysis eventually revealed HFT-related glitches as the culprit, but it took months for the SEC to analyze the data, arrive at answers, and issue a report.

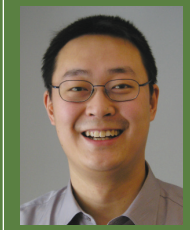
"Fifteen years ago, trade was done by humans," says Mao Ye, assistant professor of finance at the University of Illinois at Urbana-Champaign, "and you didn't need supercomputing to understand and regulate the markets. Now the players in the trading game are superfast computers. To study them you need the same power. The size of trading data has increased exponentially, and the raw data of a day can be as large as 10 gigabytes.

To address the data problem and a number of other questions related to HFT, Ye and colleagues at Illinois and Cornell University 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). Likewise, SDSC's Robert Sinkovits applied a combination of techniques to optimize Ye's code in porting it to Gordon, obtaining a 70-fold speedup in performance.

In a study they reported in July 2011, Ye, Chen Yao of Illinois, and Maureen O'Hara of Cornell processed prodigious quantities of NASDAQ historical market data — two years of trading — to look at how non-reporting of trades under 100 shares may skew perceptions of the market. Their paper, "What's Not There: The Odd-Lot Bias in TAQ Data," was featured in news reports, including *Business Week* and *Bloomberg BusinessWeek,* and has aroused policy debate. In September, as a result, the Financial Industry Regulatory Authority, which oversees the securities exchanges, reported plans to reconsider the odd-lots policy, and a vote is expected in November.

In more recent work, Ye and Illinois 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, such as a speed limit on orders or a fee for order cancellation, 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.

"Without XSEDE and shared memory," says Ye, "we wouldn't be able to effectively study these large amounts of data produced by high-frequency trading."



Mao Ye, University of Illinois at Urbana-Champaign

Researcher: Mao Ye, University of Illinois at Urbana-Champaign

Funding: Bureau of Economic and Business Research and Research Board, University of Illinois.

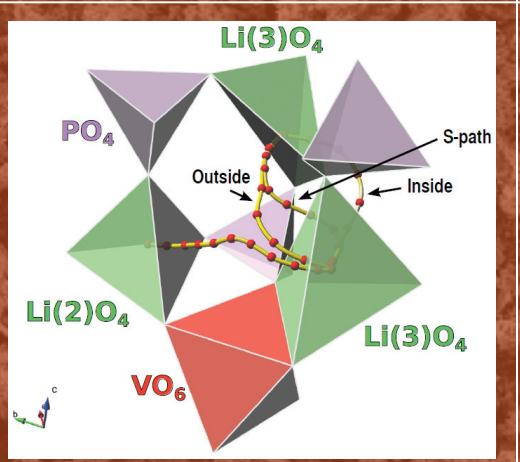
Grant no.: SES120001

For more information: http://www.sciencedirect.com/science/article/pii/S0304405X11000390

Story by Michael Schneider

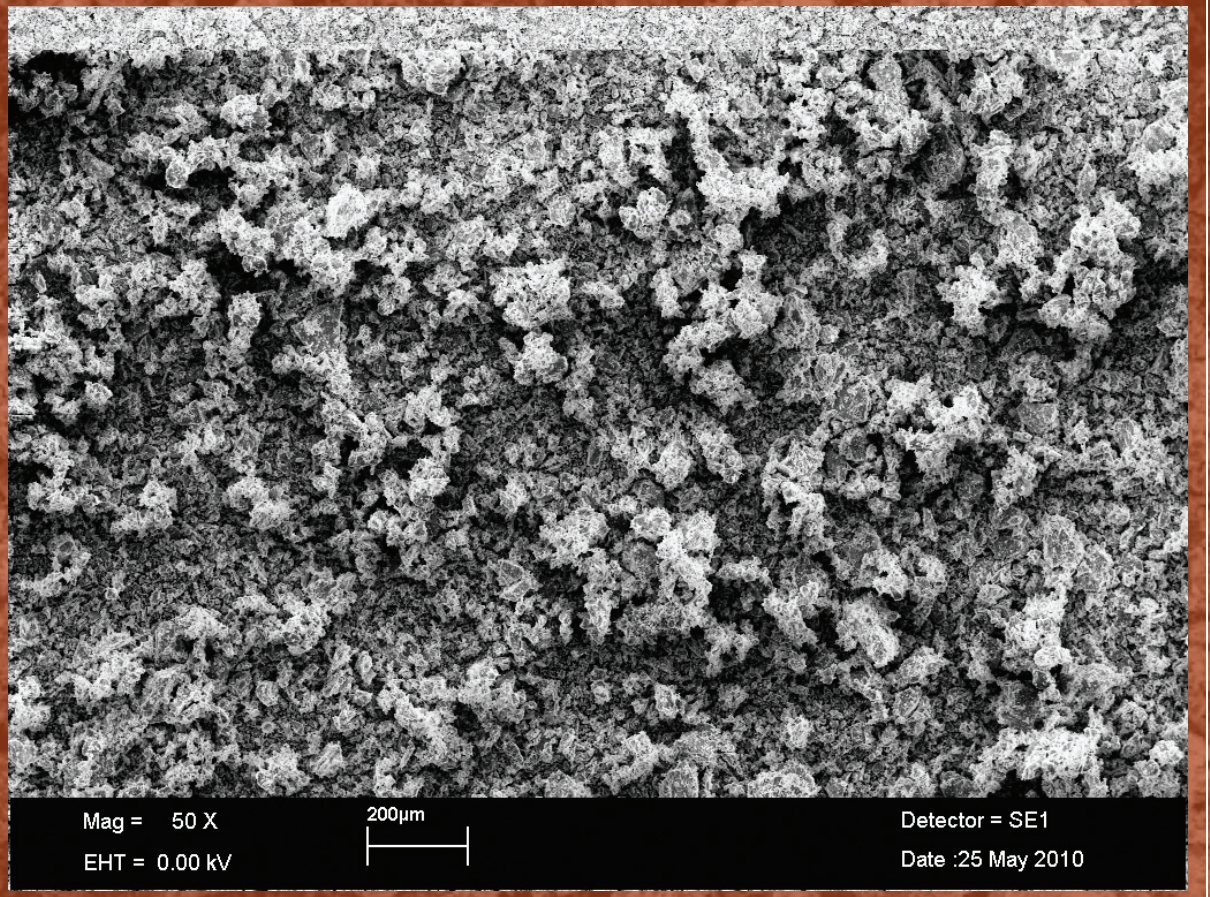
Better Batteries through Simulation

MIT researchers use Ranger supercomputer to investigate new material for high-density energy storage



The figure shows the different ways lithium (Li) might diffuse between sites in the structure. The calculations revealed that the rate-limiting step for Li diffusion in this case becomes Li migrating from the Li2 site to Li3 site.

(This figure was generated by Charles Moore using data from calculations run at TACC.)



Scanning electron microscope image showing the microstructure of the as-synthesized Li9V3(P2O7)3(PO4)2 compound.

Most of the Earth — in fact most of the universe — is made up of inorganic materials formed by geological or cosmological processes. Despite their variety, all inorganic materials are composed of a not-so-terribly-large number of inorganic compounds: 50,000 to 200,000, depending on how you count. People have been studying these materials for millennia, but less than 1 percent of their properties have been explored.

Gerbrand Ceder, professor of materials science and engineering at the Massachusetts Institute of Technology (MIT), is using some of the world's most powerful supercomputers to address this knowledge gap. He estimates that the Ranger supercomputer at the Texas Advanced Computing Center could calculate a given property for all known compounds in 80 hours, and Ranger is just one of a pantheon of powerful systems in the United States that are available to scientists as part of the XSEDE program.

A new day is dawning in materials science, an interdisciplinary field that investigates the relationship between the structure of materials at atomic or molecular scales and their real-world properties. Thanks to the enormous advanced computing resources funded by the National Science Foundation, discoveries are now only a simulation away.

Ceder and his research team's computer-based investigations into new forms of lithium ion interfaces recently led to the discovery of a novel high-density cathode material with improved characteristics over the leading alternatives. The simulations also provided insights into the pathways by which lithium and other elements cycle through batteries. The findings were reported in the *Journal of The Electrochemical Society* in March 2012.

"The nice thing about computations is that they tell you what may be possible," says Ceder. "Having them available really helps you organize your work and thinking."

Ceder hopes this incredible rate of discovery and rapid prototyping will lead to swift advances in the areas that matter most to people, among them sustainable energy production. Breakthroughs in energy density are required if the nation hopes to achieve a widespread use of electric cars. Similarly, stabilizing the country's electric power grid using batteries would require materials that can run for decades — a far cry from today's technologies.

In their paper, Ceder and his team describe the creation of lithium pyrophosphate ($Li_9M_3(P_2O_7)_3(PO_4)_2$) — a new material that never existed before — by means of artificial intelligence calculations performed on local clusters at MIT. The algorithm altered existing materials to produce new structures that are stable and display new and improved attributes.

"I've been in this field for 20 years, and I never would have dreamed of this material," Ceder says.

The researchers turned to the Ranger supercomputer to perform diffusion calculations for the new material. The simulations led the scientists to understand why the material worked better than its less-complex relatives and how it can be further improved. The researchers then synthesized and tested the material in the lab. It produced excellent energy density (the amount of power stored in a given system per unit volume), matching the simulations.

"The XSEDE resources helped us to evaluate lithium diffusivity in our predicted (and subsequently experimentally realized) lithium pyrophosphate material, for use as battery cathodes," said Anubhav Jain, co-author of the paper and discoverer of the new compound. "Future efforts to improve diffusion in this material might focus on tuning layer spacing as was previously realized for layered metal oxides."

In the past, it took an average of 18 years from the discovery of a new material to its commercialization. "An excruciating amount of time," Ceder notes. Rapid computational search and exploration has changed all that. In the last 18 months, the startup company Ceder co-founded, Pellion Technologies, has patented more insertion cathodes for experimental magnesium batteries than have been invented for lithium ion batteries in the last 25 years.

The road from a predicted material to a breakthrough product is still long and winding, but Ceder is hopeful about the prospect of facing these challenges.

"It's so exciting that we can actually design materials in a computer and go and make them," says Ceder. "For a lot of fields of engineering that seems so obvious: one can design a building and build it. But in materials science, that has almost never been done. We're entering an era of designer materials."

Gerbrand Ceder, Massachusetts Institute of Technology

Researcher: Gerbrand Ceder, Massachusetts Institute of Technology (MIT)

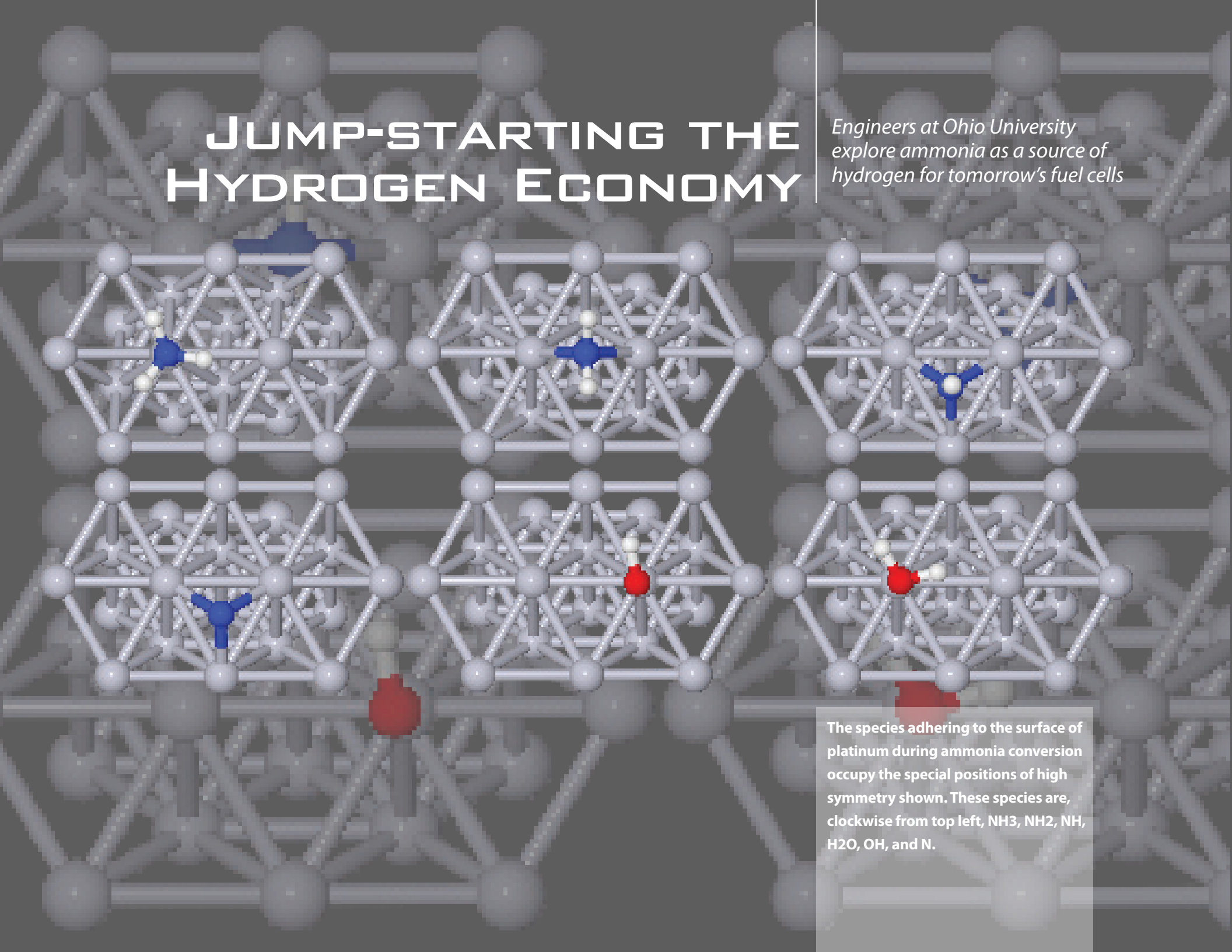
Grant no.: 0941043

For more information: www.materialsproject.org

Story by Aaron Dubrow

Jump-starting the Hydrogen Economy

Engineers at Ohio University explore ammonia as a source of hydrogen for tomorrow's fuel cells



The species adhering to the surface of platinum during ammonia conversion occupy the special positions of high symmetry shown. These species are, clockwise from top left, NH3, NH2, NH, H2O, OH, and N.

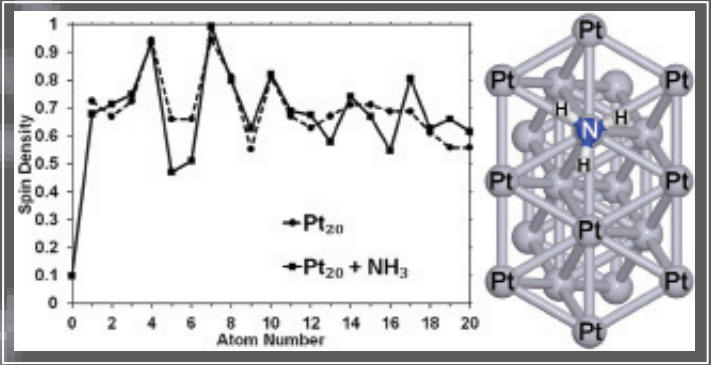
For the fuel cell industry, ammonia has all kinds of things going for it. It's abundant — the U.S. Geological Survey reports that more than 130 million metric tons were produced worldwide in 2009. The pipelines and trucks to ship ammonia around the country are already in place. It releases nothing but nitrogen and water vapor into the atmosphere when used for fuel cells. And, according to the Iowa Energy Center, it's at least as safe as gasoline when used as a transportation fuel.

But there's a downside. It requires a lot of heat to break the hydrogen — which is what most fuel cells use to generate energy — out of the ammonia, and by-products from the process can foul the fuel cells and reduce their efficiency.

Using the recently retired Ember system at the National Center for Supercomputing Applications and Blacklight at the Pittsburgh Supercomputing Center, an Ohio University team led by Professor Gerardine Botte is jump-starting the prospects of ammonia-based fuel cells. The team's results were published in 2012 in *Computational and Theoretical Chemistry*. Their approach to breaking the ammonia down for use in fuel cells also has been licensed by E3 Clean Technologies for possible use in cleaning up wastewater.

In a fuel cell, hydrogen atoms are split into protons and electrons by using a small electric current and a catalyst to drive the chemical reaction. The protons pass through a selective membrane, while the electrons are forced through an external circuit to generate electricity.

In the case of an ammonia-based fuel cell, the hydrogen is created by "cracking" the ammonia into its component parts. Older designs for cracking ammonia were inefficient, but using Botte's patented method, 1.55 watt-hours of electric energy yields a gram of hydrogen from ammonia. That hydrogen can then produce 33 watt-hours of electricity from a fuel cell.



Spin density plot for the Pt20 cluster and adsorbed NH3 on Pt20 cluster, showing minimal change in spin upon NH3 adsorption.

Courtesy of: Gerardine G. Botte, Center for Electrochemical Engineering Research, Ohio University.

One kilogram of hydrogen produced this way costs about 90 cents. Using water to produce that kilogram of hydrogen runs about $7. And that 90-cent kilogram of hydrogen can deliver roughly the same amount of energy as a gallon of gas.

The Botte team uses both experimental and computational methods. They build "electrolyzers" that zap ammonia into its component parts to test the efficiency of the system and how changes — such as changing the elements used to make the catalyst — impact that efficiency.

Using Ember and Blacklight, the team models the molecules that form and are then further broken down in the process of sheering hydrogen from ammonia molecules. They consider the strength of the bond between these intermediates and the platinum catalyst that drives the interaction. They also look at the orientation of those molecules as they interact with the catalyst, as well as how much and in what ways individual atoms within the intermediates are moving.

They then rank the intermediates in terms of which are most likely to adhere to the platinum, which in turn tells them which are most likely to pollute the catalyst over time and impinge on how well it works.

Understanding these features allows the team to look for a Goldilocks reaction that produces the most hydrogen while degrading the catalyst as little as possible. It also means they can begin to explore platinum alloys — platinum mixed with elements like iridium or rhodium — that might be more efficient or cost less money.

"XSEDE was important to the project, as the equations used to calculate the properties of the system are large and thus require the use of large memory for temporary storage and fast processors for timely results," says Damilola Daramola, a post-doctoral research associate working with Botte. "Consequently, supercomputing resources were essential and imperative."

The Botte research team (NOTE: Gerardine Botte is left-center in green turtleneck and silver necklace).

Researcher: Gerardine Botte, Ohio University

Funding Institution: U.S. Army Construction Engineering Research Laboratory

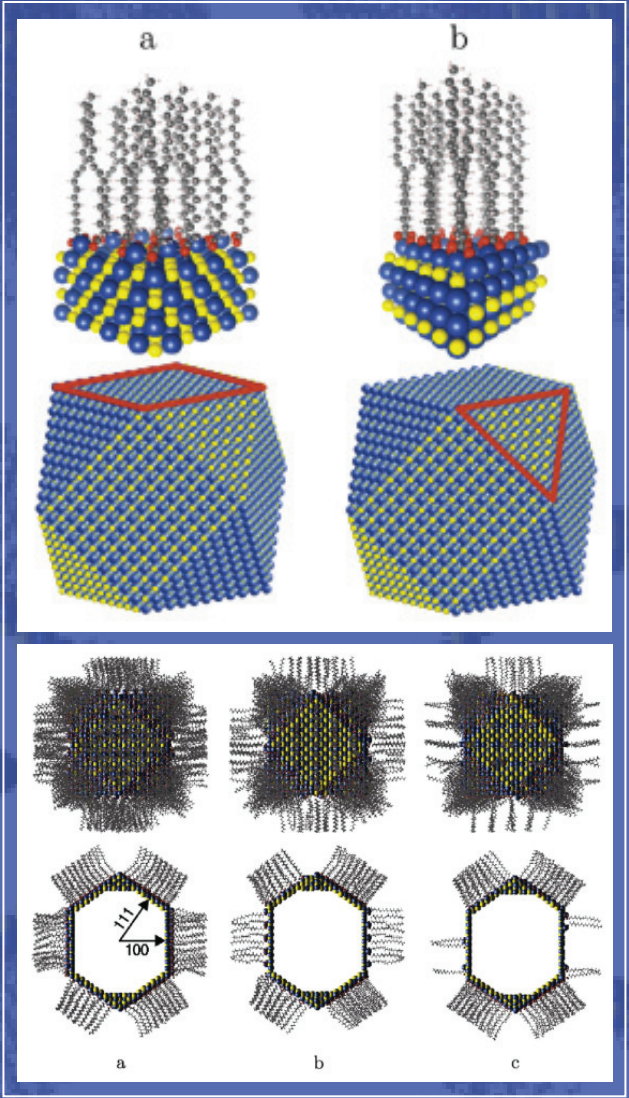
Grant no.: W9132T-09-10001

For more information: http://www.ohio.edu/ceer/

Story by J. William Bell

Design Principles for Nanoparticles

Cornell researchers use Ranger supercomputer to investigate nanocrystals for photovoltaics and catalysis



Nanoparticles are increasingly found in everyday goods, from sensors to skin care products to dirt-repelling khakis. Minute collections of atoms — typically from 1 to 100 nanometers in diameter — nanoparticles behave differently than their larger cousins, often exhibiting extreme characteristics such as stickiness, slipperiness, or hardness.

There's nothing new about nanoparticles. Artisans used them as far back as 9th century Mesopotamia to generate glittering optical effects on the surface of pots. What *is* new is our ability to produce, control, and design new nanoparticles for applications.

One of the most promising areas of nanotechnology innovation lies in new energy applications, where nanoparticles can perform nanoprocesses — such as separating hydrogen from oxygen in water or transforming photons of light into useable power — with far greater efficiency than larger molecules.

Over the last decade, researchers have developed a cookbook of recipes to produce useful nanoparticles for all sorts of industrial applications. However, a comprehensive understanding of the "design principles" behind nanotechnology continues to develop. This is largely because nanoparticles are too small and form too quickly to be captured by microscopes or other imaging devices. For that reason, one of the best ways to understand how nanoparticles form and operate is through computer models and simulations.

Richard Hennig, assistant professor of materials science and engineering at Cornell University and a 2011 recipient of the NSF CAREER award, is working to uncover the design principles at play in the formation of nanocrystals relevant to energy applications.

Using the Ranger high-performance computing system at the Texas Advanced Computing Center, as well as those at the Computation Center for Nanotechnology Innovation at Rensselaer Polytechnic Institute, Hennig and his research team showed that the concentration and location of small molecules (ligands) on the surface of lead-selenium nanoparticles cause the particles to form different shapes with different energy potentials. The results of the study were published in *ACS Nano* in February 2012.

Lead-selenium and other lead salts are a common and well-studied material used in photovoltaic cells. When the individual nanocrystals line up into ordered superstructures, they maintain the ideal band gap and electronic properties to produce electricity from the sun.

However, the nanocrystals can form a range of shapes and assemble into different superstructures that are more or less efficient. Hennig's study focused on two questions: What controls the shape of the nanocrystals, and what controls their assembly?

By altering the concentration of ligands present in his simulations of nanocrystals, Hennig and his team produced a range of shapes from octahedrons to cubes with cutoff corners.

"Experimentally, you're a little bit at a loss because we don't really know how these nanoparticles interact with each other at that scale," Hennig says, adding that simulations help determine what controls the assembly of nanoparticles into different crystal structures.

According to Hennig's research, the nanoparticles are almost never bare. Small ligands attach to their surface, like hairs on a head. Their overall concentration and specific placement (including "bald spots") appear to control the shape of the larger crystal. These results, determined by computer simulations on the XSEDE-allocated Ranger supercomputer using density functional calculations, were confirmed through laboratory experiments. By capturing the dynamics of multiple, interacting nanoparticles over time with atomic resolution, the simulations provided additional detail on how nanoparticles behave.

"The reason I decided to use XSEDE for this project, and many other projects in my group, is that the machines are state-of-the-art, the online documentation is excellent, the technical support responds fast to user problems, the batch system is easy to understand, and most software packages are already ported to the machines," says Hennig.

In their next round of simulations, the researchers are working with various experimental groups to select different ligands to place on the surface of nanoparticles. Hennig likens the work to playing with Legos.

"The nanoparticles are like your Lego pieces, and we change how these Lego pieces can stick together," adds Hennig.

Understanding the design principles that govern the formation of nanostructures and designing these structures more rationally will ultimately speed up how scientists develop materials for photovoltaic cells, bio-medical applications and catalysts for batteries (another area where Hennig's group is active).

"There's a whole world out there of different structures that you can assemble by modifying what's on the surface of the nanoparticles," observes Hennig. "The open question is: What are the useful structures and what are the structures that are just interesting? That's what the simulations can help answer."

Researcher: Richard Hennig, Cornell University

Grant no.: 1056587

For more information: http://pubs.acs.org/doi/abs/10.1021/nn3000466

Story by Aaron Dubrow

Richard Hennig, Cornell University

Training Tomorrow's Energy Scientists

Nearly 11 years ago, two young researchers, Richard Hennig and Derek Stewart, attended a materials science workshop in Santiago, Chile. Hennig, now a professor of materials science and engineering at Cornell University, and Stewart, a senior research associate at the Cornell Nanoscale Facility, stayed in contact and in 2012 organized a similar workshop, building an international bridge for the next generation of researchers.

The workshop took place at the Pontifical Catholic University of Chile in January 2012. Graduate students and early post-doctorates from the United States, Mexico, Argentina, Brazil, Colombia and Chile participated in the program, sponsored by the Pan-American Studies Institute, a joint initiative of the National Science Foundation and U.S. Department of Energy.

"The goal is to broaden participation," Hennig says. "These students will go into industry and research and will be the people making decisions 20 years down the line."

Twenty North American students and 20 South American students traveled to Santiago, Chile, to learn about the tools needed to design new materials for energy. Instructors focused on the fundamentals of computational materials science, while encouraging new ideas for solar cells, nanoparticles, and mechanical alloys.

The students also participated in hands-on tutorial sessions using the Ranger supercomputer at the Texas Advanced Computing Center, one of the powerful computing systems in XSEDE's portfolio.

Juanita Londoño-Navarro, a graduate student in materials science at the National University of Colombia, valued the training and appreciated the opportunity to continue her computations on Ranger into the next year.

"Events like this matter greatly because they build support and generate a guide for countries that do not have all the tools and machines for work and simulations," Londoño-Navarro says.

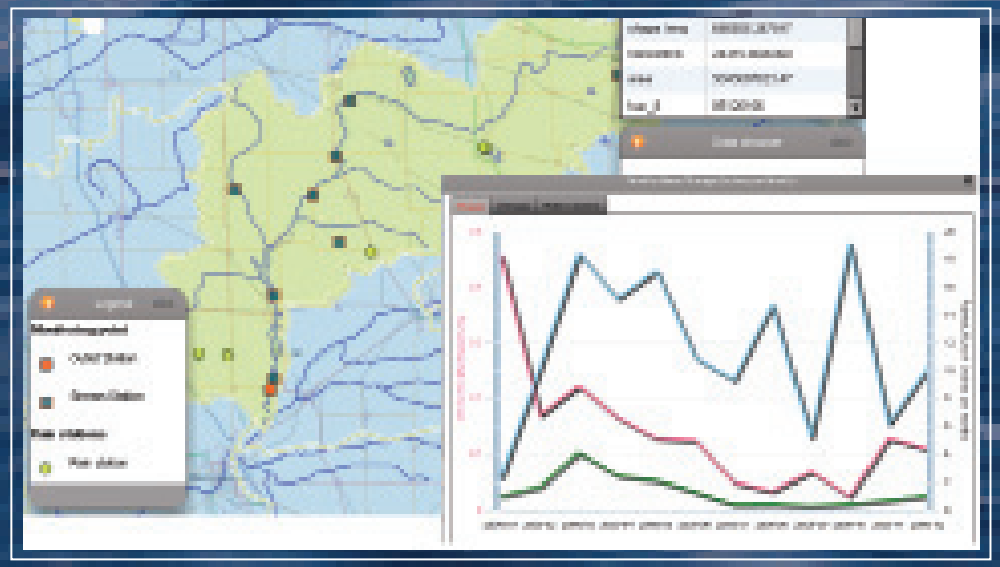
Most students were exposed to supercomputing at their universities, but 90 percent did not have access to systems of the scale provided by XSEDE.

"They may not have complete access to advanced computing resources today, but they will need them in the near future," notes Hennig. "Maybe in five to 10 years, these students will reconnect to make a workshop to show the next generation of students where they can go in life."

Lecture notes and tutorials from the event are available online on the workshop website: http://www.cnf.cornell.edu/cnf_pasi2012.html

The Hubbub about WaterHUB

With this web-based modeling tool, XSEDE computational resources flood the classroom with possibilities for studying water issues



A visualization of data on rainfall and runoff fluxes created with the Hydrology Exploration Toolkit (HET) on WaterHUB.

Water is a global issue, from availability or lack of it in the face of development and population growth to the effects of climate change, which among other things influences water-related severe weather events such as floods and droughts.

Funded by the National Science Foundation's Cyberinfrastructure Training, Education, Advancement, and Mentoring (CI-TEAM) program, the Web-based WaterHUB platform aims to support both the research and education needs of the water community. One goal is to provide a much-needed cyberinfrastructure to train and educate students, as well as the general public, on water-related issues.

A product of the WaterHUB initiative, SWATShare is a new tool — powered by XSEDE computational resources — that can be used by researchers studying water issues and by students who will become the next generation of scientists in the field. The model-sharing version of the U.S. Department of Agriculture's Soil Water Assessment Tool (SWAT) enables students to run simulations online and also to publish, share, and visualize results.

Moreover, SWATShare lets students make use of a plethora of publicly available water data rarely used in the classroom that comes from the National Climatic Data Center, the U.S. Geological Survey, and other sources.

SWAT is used in studying how land use changes — for example agricultural expansion to grow crops for biofuels or suburban growth — can influence hydrology at the watershed scale through factors such as sedimentation, stream flow, and pollution load from runoff. Likewise, the tool can be used to project the effects of climate change scenarios on watersheds.

"It is a widely used tool with a large user group," says Lan Zhao, a research programming team leader in the Scientific Solutions Group at XSEDE partner Purdue's Rosen Center for Advanced Computing, which developed SWATShare.

SWATShare made it possible for Purdue Professor Venkatesh Merwade, WaterHUB's principal investigator, to use SWAT in his computational watershed hydrology class for the first time. SWATShare offered 27 students from civil engineering, agronomy, forestry, natural resources, and other fields easy access to both the modeling tool and computational resources needed to run it effectively.

Like WaterHUB, SWATShare is built on Purdue's HUBzero platform, a ready-made open source cyberinfrastructure, which makes high-level computational tools usable in a Web browser through a friendly graphical interface. It also simplifies access to high-performance and grid computing resources such as XSEDE's. In addition to students, the Web interface can make SWAT easily accessible to teachers, researchers, decision makers, even the general public, Merwade says.

SWAT can be demanding computationally. A model run might take two or three weeks on a desktop computer. The computational resources available through XSEDE make it possible not only to get results much faster, but also to run thousands of jobs at once and examine multiple scenarios for a more complete, accurate picture. SWATShare uses the Steele cluster and the Condor distributed computing pool at Purdue, both XSEDE-allocated resources.

"We can leverage the high-performance computing resources provided by either the campus grid or national resources provided by XSEDE," Zhao says. "And with the hub, it's much easier for students to get hands-on experience."

On XSEDE, SWATShare is set up to run on a community account, which alleviates the need for students and other users to go through the process of applying for and receiving an allocation for time on XSEDE-allocated resources before they can start modeling.

HUBzero's social media-like collaboration and sharing features also make it seamless for students and other users to work together on a project and to share models of watersheds that, while they may have been created to examine one issue, can be repurposed to look at many others.

SWAT isn't the only modeling tool that will be getting the WaterHUB treatment. Among other things, Merwade says, plans call for adding standard tools for modeling precipitation runoff and for stream flow modeling.

Researcher: Venkatesh Merwade, Purdue University

More on WaterHUB: https://water-hub.org/

More on SWATShare: https://water-hub.org/swat-tool

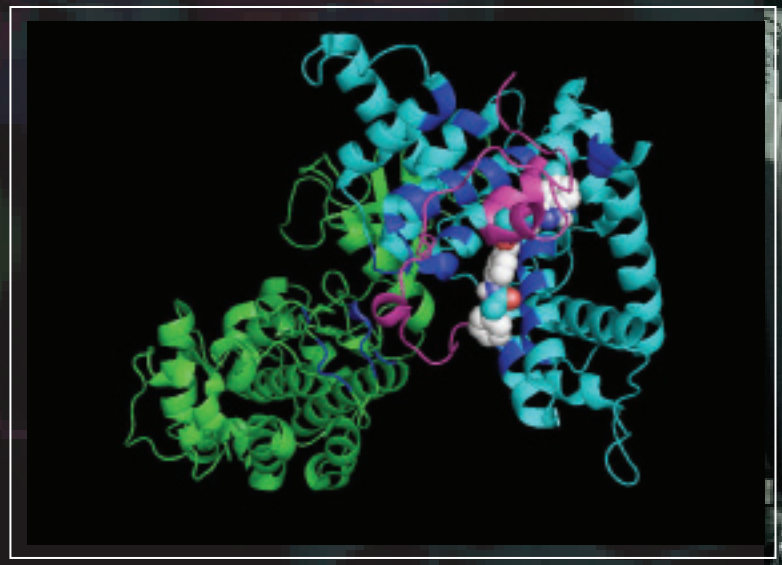
More on SWAT: http://swat.tamu.edu/

Story by Greg Kline

Venkatesh Merwade, Purdue University

Next-generation Tools for Next-generation Chemists

Undergraduates perform virtual drug screening with Ranger supercomputer



A computational model of HIV-1 Tat • human P-TEFb complex with a potential inhibitor identified by several Merrimack College students using a virtual screening approach.

Merrimack College undergraduates Daniel Laverty (back) and Dave Daniels (front) evaluate the hits from the virtual screening using PyMOL.

The drug discovery process can be arduous and expensive, sometimes requiring 15 years and hundreds of millions of dollars to find just one novel drug compound. Diseases such as cancer, HIV, diabetes, and heart disease have garnered the bulk of the research funding but tuberculosis, malaria, and other infectious diseases adversely affect the lives of many millions of people, particularly in the developing world.

Moreover, the emergence of multi-drug-resistant strains of these illnesses has highlighted the urgent need to develop novel drugs to treat them.

The sheer number of diseases that require attention necessitates a more efficient way of identifying potential drug-like compounds. One method that has gained a lot of interest in recent years is virtual screening. Virtual screening uses computational methods to identify small molecules that have a high binding affinity to a known drug target, often a protein. These become the basis for tomorrow's wonder drugs, or at least that is the hope.

One limitation to virtual screening has been the computational time required to identify a potential inhibitor. As the success rate for identifying an effective inhibitor is extremely low, a large number of compounds must be screened. XSEDE makes it possible to screen hundreds of thousands of compounds in hours, rather than months.

Virtual screening has become a valuable tool for many biotechnology and pharmaceutical companies, so educators are beginning to prepare students for the workforce by incorporating virtual screening techniques into their curriculum. Last spring at Merrimack College in Massachusetts, three professors worked together to mentor the students in Jimmy Franco's medicinal chemistry course in the virtual screening approach to drug discovery.

Students chose a protein that is suspected to be a good target for a disease they wanted to work on. They learned how to conduct a virtual screen in a lab and then learned how to conduct a virtual screen using the Ranger supercomputer from the Texas Advanced Computing Center through an XSEDE educational grant.

The students were able to compute on Ranger in part because David Toth was serving as the Campus Champion at Merrimack at the time. (He is now teaching and serving as Campus Champion at the University of Mary Washington.) Campus Champions act as a local source of knowledge about advanced computing in their communities. They provide knowledge and assistance that empowers campus researchers, educators, and students to use XSEDE to advance scientific discovery.

In this capacity, Toth was able to get an Educational Allocation on XSEDE that allowed the students to use one of the nation's most powerful supercomputers.
This activity exposed students to massive computing resources and also showed them a way of conducting science that they did not know existed. By using computational techniques, students were able to screen thousands of compounds in a short period of time, which would not be feasible in the laboratory. Compounds with high binding affinities were subsequently visualized to determine if the predicted binding orientation would inhibit the protein's function.

Feedback from the students was extremely positive, and many of them continued working on their projects even after the semester ended. Along with better preparing students for careers in the pharmaceutical and biotechnology industries, virtual screening lends itself very well to teaching many vital aspects of biology, biochemistry, and chemistry.

As instance, virtual screening relies on thermodynamics to determine the binding affinity of the potential inhibitors. While students often fail to understand the importance of thermodynamic concepts in the abstract, the drug discovery project allowed them to see the importance of thermodynamically favorable interactions in action, the professors said, as they can mean the difference between curing a disease or a tragic outcome.

Since the docking program showed the predicted binding conformation, students could also visually investigate how the inhibitor and the target protein interact.
This gave the students valuable experience visualizing biological molecules in three dimensions, as well as reinforcing the intimate relationship between structure and function.

Lastly, the project demonstrated the importance of interdisciplinary collaboration between biology, chemistry and computer science.

"Increasingly, these fields are working together to make great discoveries," Toth says. "We made sure our students understood the opportunities that are and will be available to interdisciplinary scientists who can master the application of advanced computing."

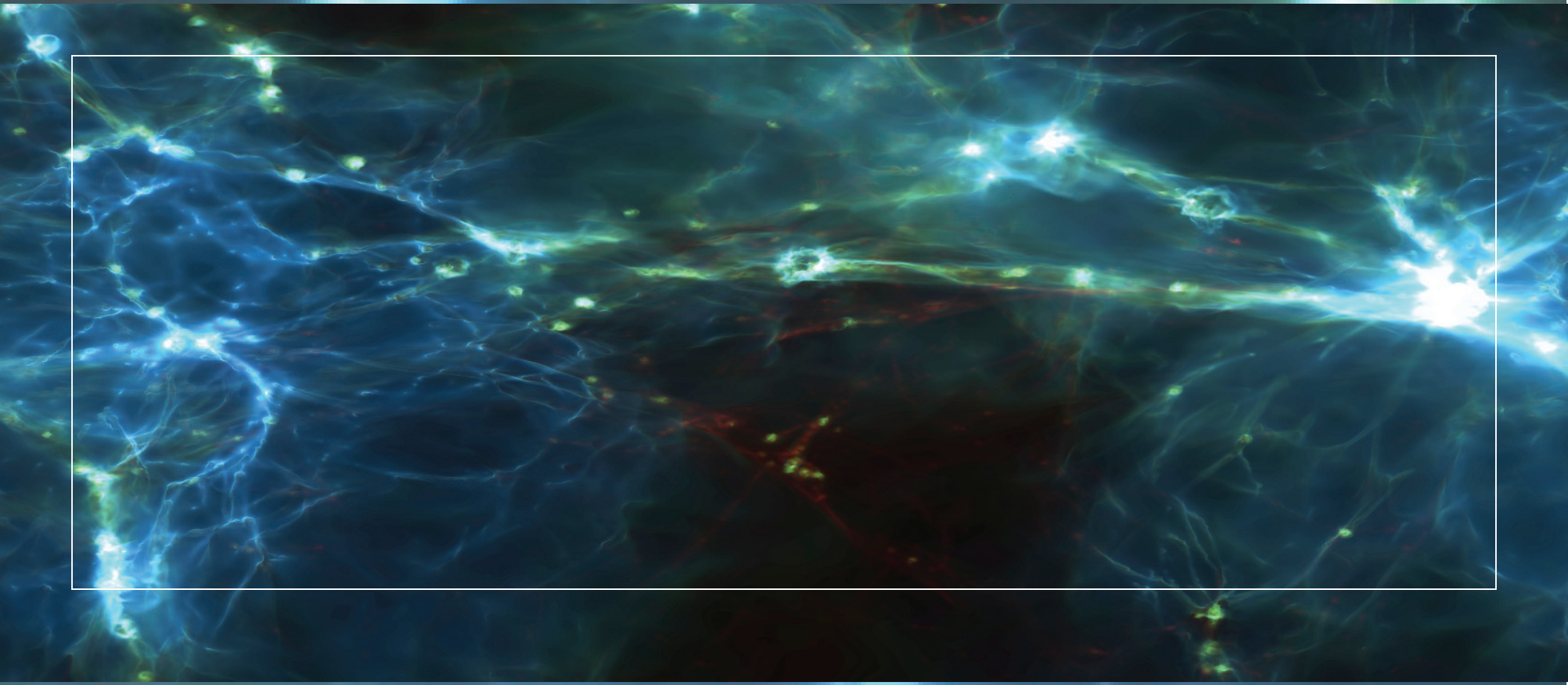
Researcher: Jimmy Franco, Merrimack College

Grant no: MCB120071

Story by
Jimmy Franco - Merrimack College
David Toth - University of Mary Washington
Charlotte Berkes - Merrimack College

Pop Star Pressures

Researchers show that radiation pressure is key to Pop II and Pop III star formations



This volume rendering (where red and blue indicates hot and cold) depicts a region 110,000 lightyears across — approximately the size of the disk of the Milky Way — 800 million years after the Big Bang.

Courtesy of John H. Wise, Georgia Institute of Technology

A team of astrophysicists from across the country has generated what they believe to be the first simulations to include the effects of radiation pressure in regulating the formation of stars in high-redshift dwarf galaxies, or some of the first galaxies to form in the universe.

These detailed simulations of what are called Pop II (oldest observed) and Pop III (oldest unobserved) stars show that radiation pressure regulates star formation in dwarf galaxies, in addition to the effects of supernovae gases and heating from starlight.

"The inclusion of radiation pressure calculations in galaxy formation simulations is a crucial element to forming realistic stellar populations and avoiding the overcooling problem, even in high-resolution simulations which capture star forming regions with many computational elements," said John H. Wise, an assistant professor of physics at the Georgia Institute of Technology and principal investigator of a new paper, called "Birth of a Galaxy. II. The Role of Radiation Pressure," and to be published in late 2012 in the *Monthly Notices of the Royal Astronomical Society*.

Researchers used the Trestles computer cluster at the San Diego Supercomputer Center (SDSC) at the University of California, San Diego, and Kraken at the National Institute for Computational Sciences (NICS) at the University of Tennessee Knoxville to conduct three distinct cosmological radiation hydrodynamics simulations to demonstrate the role radiation pressure plays as stars form in dwarf galaxies.

The researchers progressively added more physics to each simulation to better understand the impact of each process. The first model, used as a reference, considered primordial chemistry in Pop II and III star formation, while radiative cooling from fine-structure transitions in metals was added to the second one using ENZO 2.0 computer code, which accommodates the dynamic range of spatial and temporal steps needed to model the volume of the universe.

"In our final and most realistic model, we added an H2-dissociating radiation background and momentum input from stellar radiation," said Michael Norman, with UC San Diego's Center for Astrophysics and Space Sciences. "To our knowledge, this is the first cosmological galaxy simulation that has included the effects of radiation pressure that is computed from the radiative transfer equation."

Using XSEDE-allocated and other resources, the researchers found that radiation pressure indeed plays a key role in regulating star formation in high-redshift dwarf galaxies.

The researchers plan to apply their numerical methods and findings to simulating galaxies that are currently observed in the Hubble Ultra Deep Field, a small region of space consisting of about 10,000 galaxies that was composited by Hubble Space Telescope data accumulated over a period of almost four months in 2003 and 2004. It is still considered to be the deepest image of the universe ever taken.

John H. Wise, Georgia Institute of Technology

Researcher: John H. Wise, Georgia Institute of Technology

Grant no.: NSF-OCI-1048505, NSF-AST-0807312 and NSF-AST-1109243; NASA 120-6370, NASA -ATFP NNX08AH26G, and NASA/NCCS SMD-11-2258

For more information:
http://www.wiley.com/WileyCDA/WileyTitle/productCd-MNR.html

Story by Jan Zverina

XSEDE
Extreme Science and Engineering Discovery Environment

XSEDE is supported by the National Science Foundation.

xsede.org

XSEDE Leadership Team

John Towns, project director
NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS
University of Illinois at Urbana-Champaign

Tim Cockerill, associate project director
NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS
University of Illinois at Urbana-Champaign

John Boisseau, director of user services
TEXAS ADVANCED COMPUTING CENTER
The University of Texas at Austin

Chris Hempel, deputy director, user services
TEXAS ADVANCED COMPUTING CENTER
The University of Texas at Austin

Nancy Wilkins-Diehr, director of Extended Collaborative Support Services-communities
SAN DIEGO SUPERCOMPUTER CENTER
University of California, San Diego

Ralph Roskies, director of Extended Collaborative Support Services-projects
PITTSBURGH SUPERCOMPUTING CENTER
Carnegie Mellon University/University of Pittsburgh

Sergiu Sanielevici, deputy director, Extended Collaborative Support Services-projects
PITTSBURGH SUPERCOMPUTING CENTER
Carnegie Mellon University/University of Pittsburgh

Greg Peterson, director of operations
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
University of Tennessee Knoxville

Victor Hazlewood, deputy director of operations
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
University of Tennessee Knoxville

Kathlyn Boudwin, manager, project management and reporting
OAK RIDGE NATIONAL LABORATORY

David Lifka, coordinator, architecture and design
CORNELL UNIVERSITY

Ian Foster, architect, architecture and design
ARGONNE NATIONAL LABORATORY

Andrew Grimshaw, architect, architecture and design
UNIVERSITY OF VIRGINIA

JP Navarro, manager, software development and integration
ARGONNE NATIONAL LABORATORY

Janet Brown, manager, systems and software engineering
PITTSBURGH SUPERCOMPUTING CENTER
Carnegie Mellon University/University of Pittsburgh

Scott Lathrop, director, education and outreach
SHODOR EDUCATION FOUNDATION

Carol Song, chair, Service Provider Forum
PURDUE UNIVERSITY

David Hancock, vice chair, Service Provider Forum
INDIANA UNIVERSITY

Tom Cheatham, chair, user advisory committee
UNIVERSITY OF UTAH

Steven Gordon, manager, education
OHIO SUPERCOMPUTER CENTER
The Ohio State University

Laura McGinnis, manager, outreach
PITTSBURGH SUPERCOMPUTING CENTER
Carnegie Mellon University/University of Pittsburgh

Dan Stanzione, manager, training
TEXAS ADVANCED COMPUTING CENTER
The University of Texas at Austin

External Relations Team

Trish Barker
NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS
University of Illinois at Urbana-Champaign

Aaron Dubrow
TEXAS ADVANCED COMPUTING CENTER
The University of Texas at Austin

Jim Ferguson
NATIONAL INSTITUTE FOR COMPUTATIONAL SCIENCES
University of Tennessee Knoxville

Warren Froelich
SAN DIEGO SUPERCOMPUTER CENTER
University of California, San Diego

Greg Kline
INFORMATION TECHNOLOGY AT PURDUE
THE ROSEN CENTER FOR ADVANCED COMPUTING
Purdue University

Michael Schneider
PITTSBURGH SUPERCOMPUTING CENTER
Carnegie Mellon University/University of Pittsburgh

Faith Singer-Villalobos
TEXAS ADVANCED COMPUTING CENTER
The University of Texas at Austin

Jan Zverina
SAN DIEGO SUPERCOMPUTER CENTER
University of California, San Diego

Susan McKenna
NATIONAL CENTER FOR SUPERCOMPUTING APPLICATIONS
University of Illinois at Urbana-Champaign

CarltonBruettDesign