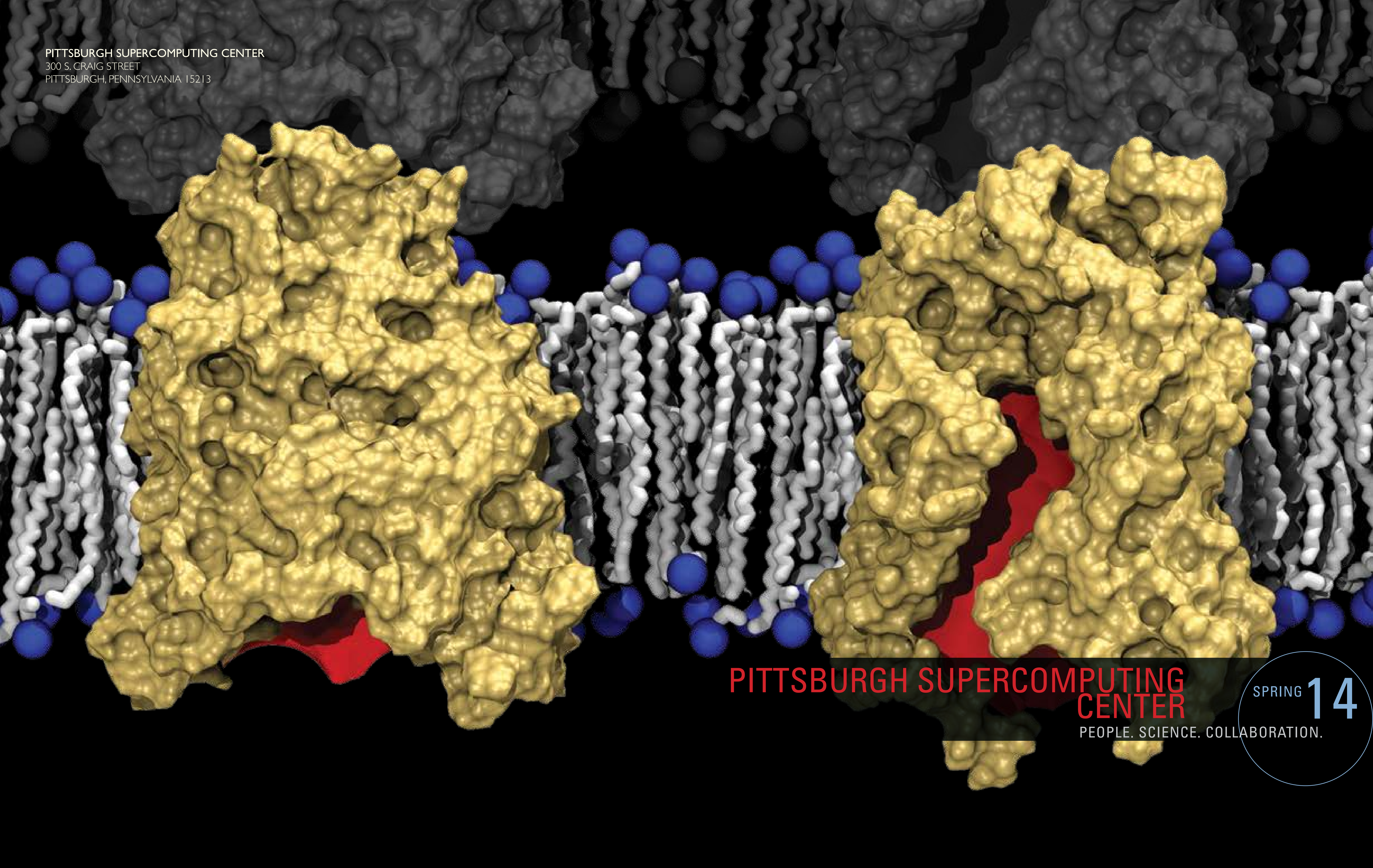
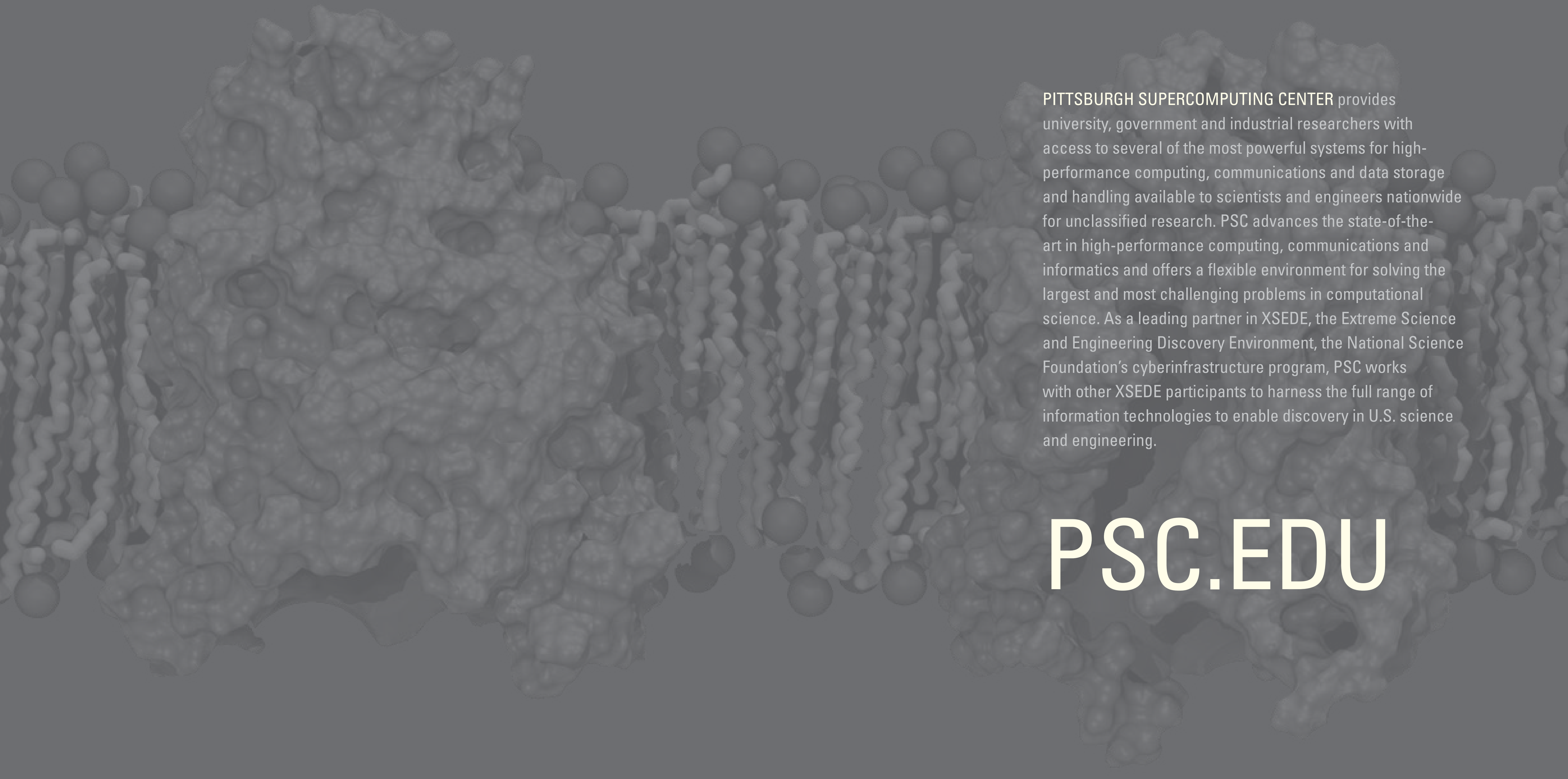


PITTSBURGH SUPERCOMPUTING CENTER
300 S. CRAIG STREET
PITTSBURGH, PENNSYLVANIA 15213

PITTSBURGH SUPERCOMPUTING CENTER

SPRING 14

PEOPLE. SCIENCE. COLLABORATION.

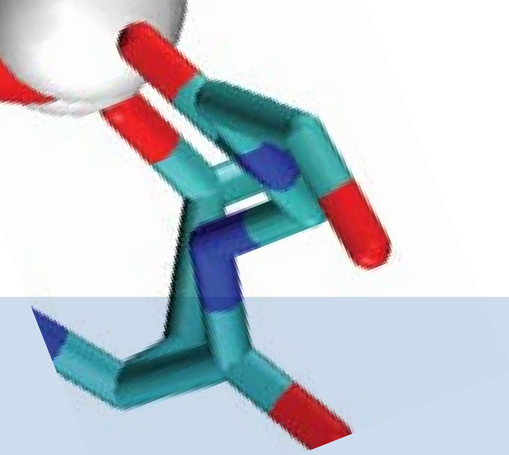


PITTSBURGH SUPERCOMPUTING CENTER provides university, government and industrial researchers with access to several of the most powerful systems for high-performance computing, communications and data storage and handling available to scientists and engineers nationwide for unclassified research. PSC 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.

PSC.EDU

SPRING 2014

CONTENTS



FROM THE **DIRECTORS**

Welcome once again to Pittsburgh Supercomputing Center's biannual report, featuring the pathbreaking projects being pursued by our users, our staff and our students. We're very pleased in this issue to cover a broad spectrum of work, including research, technical milestones and ongoing programs to train new generations of high performance computing (HPC)-savvy scientists and engineers.

The work described in the following pages ranges from HPC staples such as cosmology (*p.* 10) and molecular dynamics simulations (*pp. 18* and 24) to fields of study that are new to HPC, such as the humanities (p. *6*), public health and financial services (p. *26*). Many of these applications fit under an over-arching category, and technical challenge, of HPC: "Big Data."

While Big Data has a number of definitions, one critical aspect is that researchers are dealing with amounts of data so vast that retrieving specific information in a practical timeframe is difficult if not impossible with current infrastructure and technologies. PSC is involved in a number of projects intended to broaden and deepen our capabilities to extract information from the ballooning data being generated in science—and in many other human endeavors—more quickly and efficiently.

Since our last report PSC received a $7.6-million National Science Foundation grant to design and build the Data Exacell (DXC, *p. 14*). A new pilot system pursued with select scientific collaborators, DXC will identify optimal methods and hardware to securely and accessibly store extremely large-scale data with specialized analytics optimized for their study. Analytic resources will include Blacklight, still the largest shared-memory supercomputer available to researchers, and Sherlock, our new graph analytics supercomputer. (For a list of our supercomputing resources, see "PSC's Supercomputing Resources".)

Communication between resources—networking—is a major component of any Big Data system. PSC's renowned networking group has been producing new software and configuring hardware systems that speed, widen and prioritize network connections to avoid bottlenecks and support the largest data users (*p. 16*). Our efforts in this field are second to none in the HPC community.

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

Finally, we are pleased to learn that our national HPC colleagues have noticed our Big Data efforts (*p. 26*). This year we garnered two Reader's Choice and two Editor's Choice awards from *HPCwire* magazine (a leading trade publication for the HPC community) recognizing our efforts in genomics, public health, stock market analysis and Big Data analytics. *HPCwire* singled us out for additional recognition by naming two of our public health projects as "Top Supercomputing-Led Discoveries of 2013."

We hope you find the following accounts of our efforts over the past 6 months to be enjoyable and edifying. We would like to hear any feedback you had, on our work or this publication. You can send any comments or suggestions via our feedback page at *psc.edu/index.php/feedback*. You can also contribute to PSC's nonprofit, academic mission at psc.edu/donate.

Ralph Roskies

PSC's Supercomputing Resources

Blacklight is an SGI Altix® UV1000 shared-memory system with 4,096 cores and 32 terabytes of shared memory. It has the largest shared memory of any system available to the research community.

Sherlock is a YarcData Urika™ (Universal RDF Integration Knowledge Appliance) data appliance with PSC enhancements. It enables large-scale, rapid graph analytics through massive multithreading, a shared address space, sophisticated memory optimizations, a productive user environment and support for heterogeneous applications.

Anton is a special-purpose supercomputer for molecular dynamics simulations, designed and built by D. E. Shaw Research (DESRES). An Anton machine was provided without cost by DESRES for non-commercial use by the national biomedical community, and the machine is hosted by the National Resource for Biomedical Supercomputing at PSC. *Operational funding is provided by the NIH via grant P41GM103712-S1 to the National Center for Multiscale Modeling of Biological Systems.*

The **Data Supercell** is a low-cost, high-bandwidth, low-latency, high-capacity and high-reliability data management and archival system.

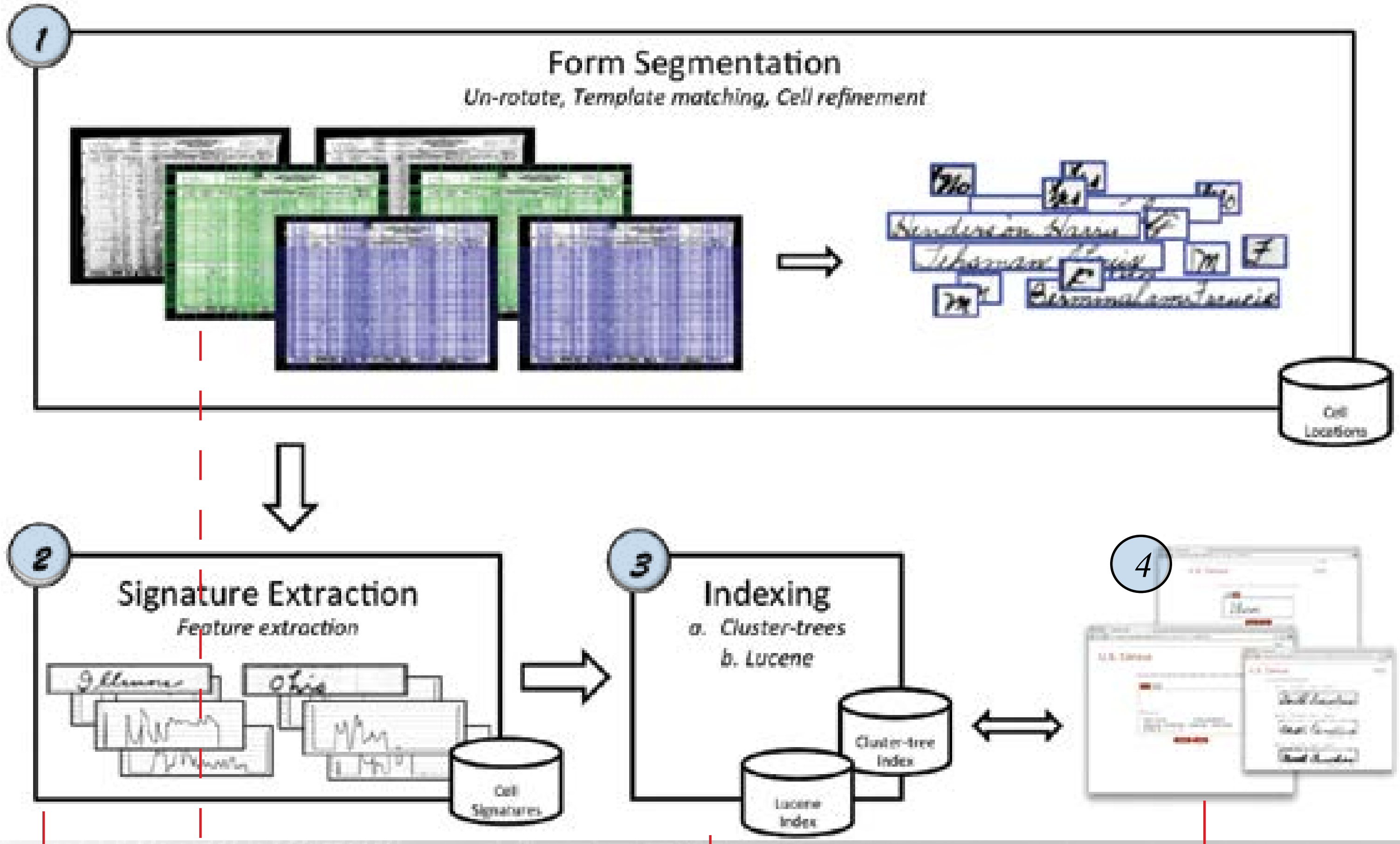
Breaking Out of the Digital Document Graveyard

Blacklight Used to Extract Meaning from Cursive Script, Allowing Scanned Documents to be Searched

In 1973, a fire broke out at the St. Louis National Personnel Records Center, destroying 16 to 18 million military service records from 1912 to 1964. If these records had been digitized they'd have been safe, but not necessarily any more accessible.

Scanned PDF images, the low-cost, high-speed method for digitizing images, can be duplicated and stored in many places. But you can't find anything in them, except by a human being searching through the handwritten text by eye. And the 1940 U.S. Census, for example, consists of 3.6 million PDF images.

How Blacklight Learned to Read Script



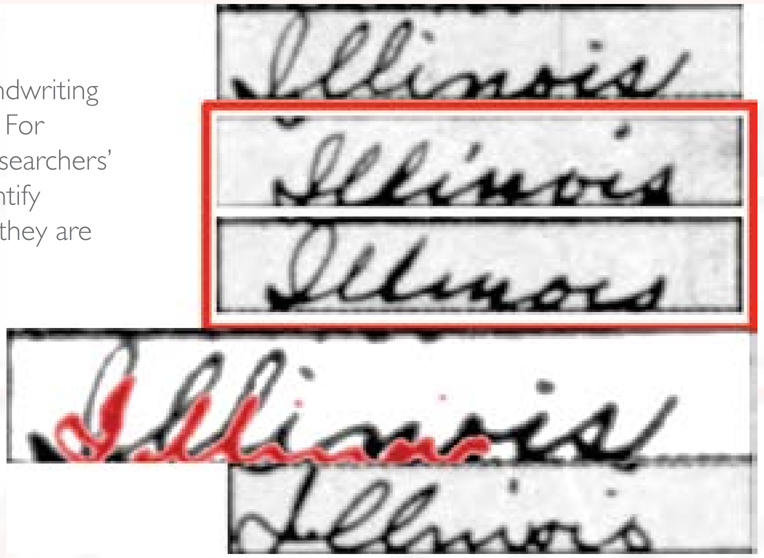
Step 1: Correct rotations and reduce smudges and bleeds to produce a spreadsheet-like document with identifiable cells.

Step 2: Create a statistical picture of the contents of each cell.

Step 3: Classify the contents' likely meaning based on these statistical pictures—"index" them.

Step 4: Users can then search for specific entries, picking the "right" answers and helping the system correct itself.

Recognizing the same word in different handwriting can be a challenge even to human readers. For computers, it was nearly impossible. The researchers' statistical method allowed Blacklight to identify cursive-script words based on how similar they are to each other.



Commercial services like Ancestry.com employ thousands of human workers who manually extract the meaning of a small, profitable subset of these images so they can be searched by computer, says Kenton McHenry of the National Center for Supercomputing Applications (NCSA), University of Illinois at Urbana-Champaign. "But government agencies just don't have the resources" necessary to make most of them accessible in this way. The danger is that scanned documents may become a digital graveyard for many historically, culturally and scientifically valuable old documents—easy to find, impossible to resuscitate.

Liana Diesendruck, an NCSA research programmer, McHenry and colleagues in his lab have used PSC's Blacklight supercomputer to begin cracking this formidable problem. Using the 1940 Census as a test case, Blacklight's architecture has allowed them to create a framework for automatically extracting the meaning from these images—in essence, teaching the machines to read cursive script.

TEACHING MACHINES TO READ

"Before we could even think about extracting information, we had to do a lot of image processing," says Diesendruck. Misalignments, smudges and tears in the paper records had to be cleaned up first. But the difficulty of that process paled compared with the task of getting the computer to understand the handwritten text.

It's relatively simple to get a computer to understand text that is typed in electronically. It knows that an "a" is an "a," and that the word "address" means what it does. But early Census entries were made by many human workers, with different handwriting and styles of cursive script. These entries can be difficult for humans to read, let alone machines.

Having the computer deconstruct each hand-written word, letter by letter, is impossible with today's technology. Instead, the investigators made the computer analyze the words statistically rather than trying to read them. Factors like the height of capital "I"s, the width of a loop in a cursive "d" and by how many degrees the letters slant from the vertical all go into a 30-dimensional vector—a mathematical description consisting of 30 measurements. This description helps the machine match words it knows with ones it doesn't.

PSC's Blacklight proved ideal for the task, McHenry says. Part of the computational problem consists of crunching data from different, largely independent entries as quickly as possible. Blacklight, while not as massively parallel as some supercomputers, has thousands of processors to do that job. More importantly, Blacklight's best-in-class shared memory allowed them easily to store the relatively massive data their system had extracted from the Census collection—a 30-dimensional vector for each word in each entry. This allowed the calculations to proceed without many return trips to the disk. Eliminating this lag to retrieve data made the calculations run far faster than possible on other supercomputers.

"GOOD ENOUGH" ACCURACY

On average, the system accurately identified words despite the idiosyncrasies of the handwriting. Of course, that "on average" is just what it means. Sometimes the machine is correct, sometimes it isn't. The idea is quickly to produce a "good enough" list of 10 or 20 entries that may match a person's query rather than taking far longer to try to make it exact.

"We get some results that aren't very good," Diesendruck says. "But the user clicks on the ones he or she is looking for. It isn't perfect, but instead of looking through thousands of entries you're looking at 10 or 20 results."

Search engines like Google have made Web users very demanding in terms of how much time a search takes. But while they expect fast, they don't expect extreme precision. They don't tend to mind scanning short lists of possible answers to their query. So the script search technology is similar to what people are used to seeing on the Web, making it more likely to be accepted by end users.

There's another virtue to how the system works, McHenry points out. "We store what they said was correct," using the human searcher's choice to identify the right answers and further improve the system. Such "crowd sourcing" allows the investigators to combine the best features of machine and human intelligence to improve the output of the system. "It's a hybrid approach that tries to keep the human in the loop as much as possible."

Today the group is using Blacklight to carry out test searches of the 1940 Census, refining the system and preparing it for searching all sorts of handwritten records. Their work will help to keep those records alive and relevant. It will also give scholars studying those records—not just in the "hard" and social sciences, but also in the humanities—the ability to use and analyze thousands of documents rather than just a few.

Cosmic Tug of War

Large Dark Matter Halos Favor Growth of Larger Early Galaxies

> *When the first stars and galaxies begin to heat the gas to 10,000 or 20,000 degrees Kelvin, the outward pressure makes it really hard for the gas to fall into smaller dark matter halos. That means that the low-mass galaxies that would have formed in smaller dark-matter halos will not get a chance to form.*
> *—Hy Trac, Carnegie Mellon University*

WHY IT'S IMPORTANT

Few scientific questions are as fundamental, or fascinating, as the origin of the Universe. And we can see the early Universe. The farthest galaxies from us are so far away that it takes light rays about 13 billion years to reach us. Our newest telescopes are, in essence, time machines that will see the light that these galaxies created just a few hundred million years after the Big Bang.

Hy Trac of Carnegie Mellon University and Renyue Cen of Princeton University lead a team of cosmologists whose simulations on PSC's Blacklight supercomputer predict that the largest early galaxies would tend to win a cosmic tug of war in galaxy formation, making it harder for smaller ones to develop. Such predictions help the big-ticket telescopes know what phenomena to look for, making them more productive.

HOW BLACKLIGHT HELPED

Trac and Cen's team uses a two-phase simulation: First, an "N-body" simulation, including only dark matter and the force of gravity, creates a framework. Adding ordinary matter and radiation then refines a more realistic, "radiation-hydrodynamic" simulation.

The latest version of the N-body simulation, run in November 2013, required almost 10 terabytes of computer memory—roughly the amount of information in all the Library of Congress' printed books. With the largest amount of "shared memory" available to academic researchers—32 terabytes—Blacklight is ideal for such calculations.

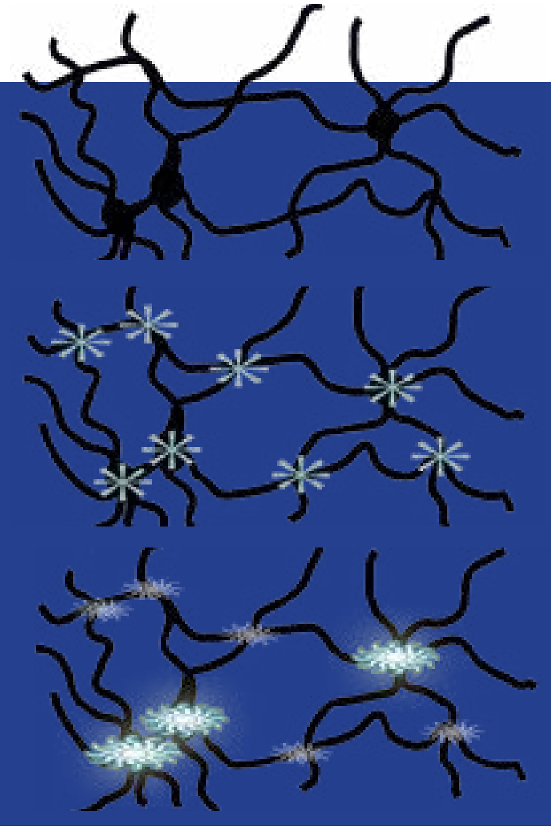
The vastness of the simulation is striking. With 70 billion particles of dark matter, it encompasses an area 300 million light years across, containing 100,000 times the mass of our Milky Way galaxy. And it's just a first step in creating a more complex radiation-hydrodynamic simulation, now under way.

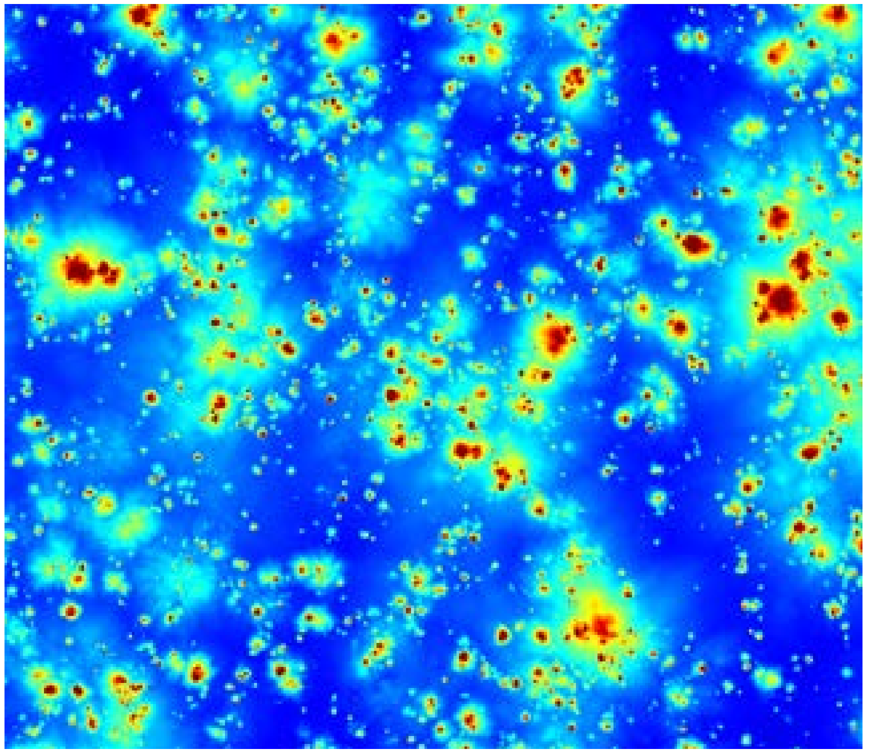
IN THE BLACKLIGHT SIMULATIONS, THE FORMATION OF THE FIRST GALAXIES FAVORS LARGER GALAXIES.

Top: Vast filaments of dark matter (black)—which account for 85 percent of the mass in the Universe—form dark-matter halos where they intersect in the early Universe. The halos' gravity attracts cosmic gas.

Middle: As the gas collects and heats, the first stars and galaxies light up (white), creating an outward burst of ionizing radiation that partly counteracts the inward pull of gravity.

Bottom: Only in the largest halos is gravity strong enough to continue collecting gas, creating large galaxies (white) in the process. Smaller halos lose the cosmic tug of war, so the smaller galaxies they would have formed (ghosted blue) are less likely to get started.





> *Our simulations will guide the next-generation observations by the James Webb Space Telescope and others. Ultimately, the dialog between simulation and observation will test the standard cosmological model and the theory of galaxy formation at the epoch of reionization.*
> *—Renyue Cen, Princeton University*

Distortion of the cosmic microwave background predicted by the team's latest radiation-hydrodynamic simulation.

Shoring Up the *Weakest* Link

© Imgur 2012

PSC Cybersecurity Group Uses Technology, User Savvy, to Guard Supercomputing Resources Nationwide

Luckily, the woman was smart.

When a waiter brought her a phone, saying her credit card company was calling her, she smelled a rat. The caller, claiming to be from her credit card company, asked for her card number "to verify her information."

She refused, which was good. The call was from a scammer. He knew her location, because he'd read the Yelp review she'd posted of the restaurant earlier that day mentioning she'd be having lunch there.

Far removed from the workings of a high-performance computing center and network? Possibly, but it does illustrate an important point. The creativity of scammers, hackers and crackers is boundless, and whether you're protecting a personal credit account or a $20-million supercomputer, you need to be on your guard. Humans will always be the weakest link in any security system.

"I've not previously heard of a scam using this technique to try to obtain information from a victim like that," says Jim Marsteller, PSC security officer. "But we're probably going to see more of that sort of thing. Being aware of and understanding how all these pieces of information can be connected and how that information can potentially be used is the best strategy for avoiding making ourselves vulnerable."

With Shane Filus, PSC information security engineer, Marsteller runs PSC's Cybersecurity Group, protecting the Pittsburgh center's systems and the larger National Science Foundation XSEDE network of computing centers from unauthorized users. Marsteller is co-lead for Cybersecurity and Incidence Response among XSEDE resources across the country, with Brandy Butler at the National Center for Supercomputing Applications at the University of Illinois. He's also co-principal investigator for the Center for Trustworthy Scientific Infrastructure, an NSF-funded effort to help researchers protect their projects and data.

FACING THE THREATS

The threat is real. Some foreign governments use cyberespionage to try to steal government, industrial and research data. "Hacktivists" may try to access, steal or corrupt data generated by scientists whose results they dislike. And organized crime has gone digital, by compromising networking and computational resources for Bitcoin mining and spam generation and delivery, as well as other unauthorized uses.

Paired with the threat posed by these bad guys is a mission that says the good guys—the scientists trying to understand the Universe and answer important practical questions—do what they do best when they can share and use information openly.

"In a business environment, Information Technology access controls are more restrictive," Marsteller says. "You know who your customers are and have greater control over the infrastructure. Our field, on the other hand, is one of the most challenging just from the aspect that we're open, we really want to foster collaboration." Balancing the need for access with the need for cybersecurity is a big priority in any academic research enterprise. To achieve that goal, the PSC group employs some of the most sophisticated tools available—and measures as simple as user education.

TOOLS OF THE TRADE

Defense against cyberintruders requires a lot of system monitoring. Such intrusion detection takes two forms: signature based and heuristic based.

"In signature-based detection, we know that a certain type of unauthorized activity has identifiable characteristics," Marsteller says. "For example, it uses a certain type of protocol and communicates on port 'X.' As the system monitors network traffic … it can then flag these signatures and say, 'This is malicious.'"

Heuristic-based detection, on the other hand, tries to identify patterns of unusual activity that stand out from normal use of a system. It then notifies administrators, who can examine the activity in more detail. Signature-based detection helps guard against known types of cyber-attack; heuristic-based detection helps catch attacks that haven't been seen before.

But users' good practice is the first and best line of defense, they both agree.

"User education is the best way to prevent intrusions," Filus says. The great majority of security incidents the group deals with stem from compromised user accounts. Users' cybersecurity "hygiene," and common sense, is therefore a key approach to keeping interlopers out. ("Cybersecurity Top 8 Dos and Don'ts")

"The more we can do to make them more aware and to get them to understand how critical they are in this whole security ecosystem, the better," Marsteller adds.

Cybersecurity Top 8 ***Dos*** and ***Don'ts***

Don't use children's birthdays, nicknames, the word "password," and the like for passwords: Did you really need to hear that again?

Do understand the level of threat: Your bank account needs greater protection than your Gmail account. So does your XSEDE account. Give the higher-value accounts unique, and more sophisticated, passwords.

Don't share: laptops, accounts, passwords.

Do use a password management tool: It lets you generate random passwords that are much harder to crack and helps to securely (and easily) manage the many accounts we have today.

Don't trust random software, particularly if it's free: Is it sharing your data with the world? And do you really need it on your work computer?

Do find a "cybersecurity buddy": Whether it's an IT caseworker or just a savvy coworker, it helps to have a second pair of eyes look at that link before you click on it.

Don't trust unexpected emails, particularly those warning you of security risks: Is the address after "@" different from the company's Web address? Are there obvious grammar errors?

Do trust your first reaction: Often we see the risk, then rationalize ourselves into ignoring it.

THINKING

$7.6-Million NSF Grant to fund the Data Exacell, PSC's Next-Generation System for Storing, Analyzing Big Data

50 to 100 thousand gigabytes of data generated by a single astronomy project on the new generation of telescopes

38 million video-hours being uploaded annually to YouTube

2.2-mile height of the stack of DVDs to store DNA sequence data being generated annually

5 million million total gigabytes of data generated worldwide annually

The term "Big Data" has become a buzzword. Like any buzzword, its definition is fairly malleable, carrying different meanings in research, technology, medicine, business and government.

One common thread, though, is that Big Data represents volumes of data that are so large that they are outgrowing the available infrastructure for handling them. In many cases, research can't be done because the tools don't yet exist for managing and analyzing the data in a reasonable amount of time. Ultimately, we need to develop both tools and an overall strategy to make Big Data fulfill its promise in fields as disparate as biomedicine, the humanities, public health, astronomy and more.

PSC is taking the next step in developing both tools and direction for harnessing Big Data. A new National Science Foundation (NSF) grant will fund a PSC project to develop a prototype Data Exacell (DXC), a next-generation system for storing, handling and analyzing vast amounts of data. The $7.6-million, four-year grant will allow PSC to design, build, test and refine DXC in collaboration with selected scientific research projects that face unique challenges in working with and analyzing Big Data.

"We are very pleased with this opportunity to continue working cooperatively to advance the state of the art based on our historical strengths in information technologies," says Subra Suresh, the president of Carnegie Mellon University.

"The Data Exacell holds promise to provide advances in a wide range of important scientific research," says Mark Nordenberg, chancellor of the University of Pittsburgh.

Big Data is a broad field that encompasses both traditional high-performance computing and also other fields of technology and of research. But these fields increasingly share a focus more on data collection and analysis—handling and understanding unprecedented amounts of data—than on computation. They also require access methods and performance beyond the capability of traditional large data stores. The DXC project will directly address these required enhancements.

"The focus of this project is Big Data storage, retrieval and analysis," says Michael Levine, PSC scientific director. "The Data Exacell prototype builds on our successful, innovative activities with a variety of data storage and analysis systems."

The core of DXC will be SLASH2, PSC's production software for managing and moving data volumes that otherwise would be unmanageable.

"What's needed is a distributed, integrated system that allows researchers to collaboratively analyze cross-domain data without the performance roadblocks that are typically associated with Big Data," says Nick Nystrom, director of strategic applications at PSC. "One result of this effort will be a robust, multifunctional system for Big Data analytics that will be ready for expansion into a large, production system."

DXC will concentrate primarily on enhancing support for data-intensive research. PSC external collaborators from a variety of fields will work closely with the center's scientists to ensure the system's applicability to existing problems and its ability to serve as a model for future systems. The collaborating fields are expected to include genomics, radio astronomy, analysis of multimedia data and other fields. (See below.)

"The Data Exacell will have a heavy focus on how the system will be used," says J. Ray Scott, PSC director of systems and operations. "We'll start with a targeted set of users who will get results but who are experienced enough to help us work through the challenges of making it production quality."

Initial DXC Partners

- National Radio Astronomy Observatory, Green Bank, WV
- Event Detection in Multimedia Project, Carnegie Mellon University
- Galaxy Genome Project, Pennsylvania State University
- Department of Biomedical Informatics, University of Pittsburgh
- World History Data Center, University of Pittsburgh

Building a 21st Century Data Highway

> “DANCES will help supercomputing users employ network resources more efficiently and effectively by scheduling and managing the movement of their data.
> —Kathy Benninger, DANCES principal investigator

PSC Advanced Networking Group Expands Networking Capabilities for Region, World

In the era of “Big Data,” the challenge of moving the vastly expanded data volumes created and needed by today’s researchers has become central. The old network—the equivalent of an overcrowded two-lane road—is giving way to a more flexible, software-defined network that manages itself and talks with users to help them work smarter. PSC has contributed to expanding and managing the hardware “lanes” and “merges” so that large-scale users can avoid the traffic jams—and just as importantly, avoid creating them.

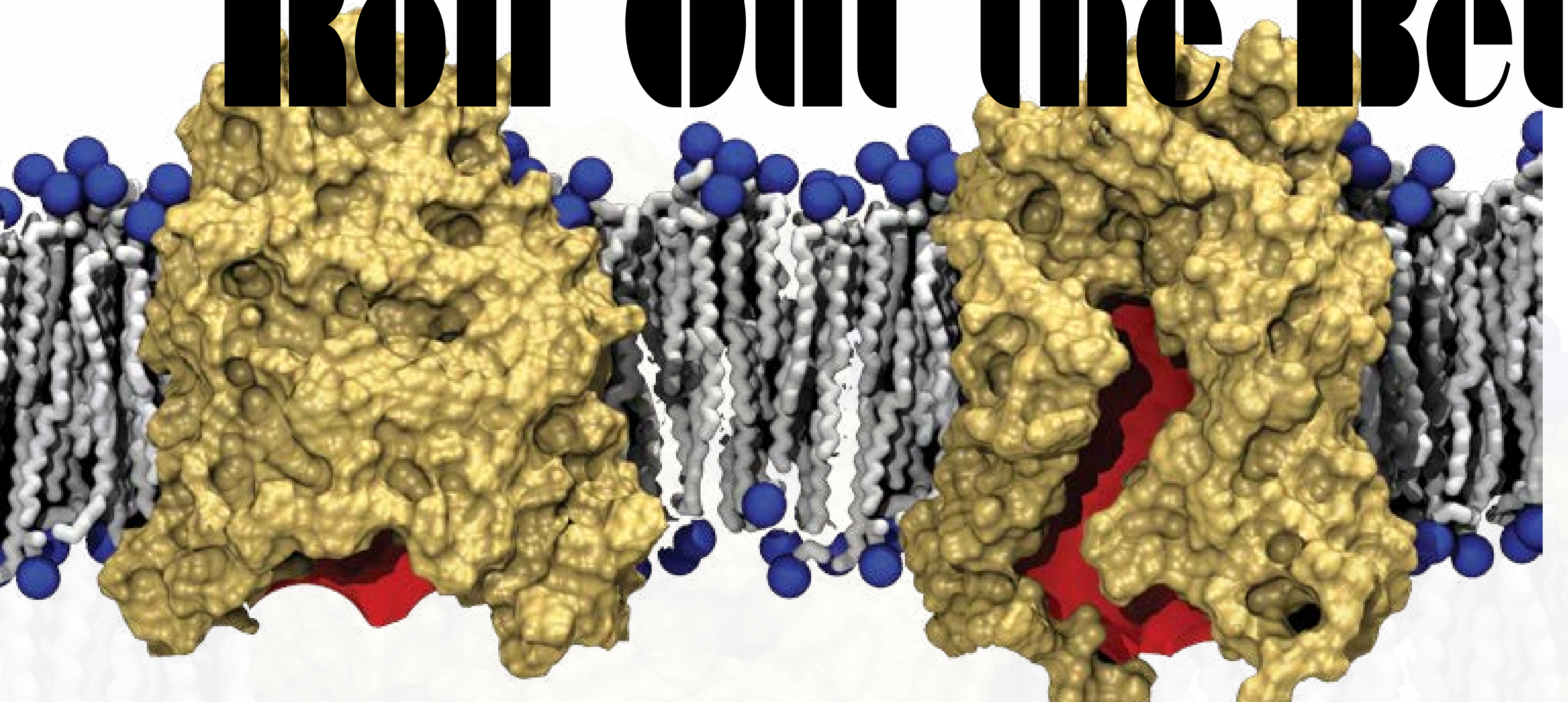
The **Web10G** collaboration between PSC and the National Center for Supercomputing Applications won an NSF grant to create a “dashboard” for networking users. Employing network data extracted by earlier Web10G work, the tool will allow nontechnical users to spot network slowdowns and report them to system administrators for repair.

Thanks to another NSF grant, PSC has “widened the highway,” activating the Pittsburgh region’s first **100-gigabit**-per-second Internet2 connection. The improved Three Rivers Optical Exchange (3ROX) connection makes it possible to move data 10 times faster than the previous fastest research and industrial connections—and 5,000 times faster than the fastest home Internet connections.

The newest NSF networking grant is for **DANCES**. A system of upgraded user-end hardware and innovative software, DANCES will allow high-volume users to schedule existing network resources. The idea is to prioritize routing and scheduling for the biggest jobs, creating virtual direct connections that avoid traffic jams.

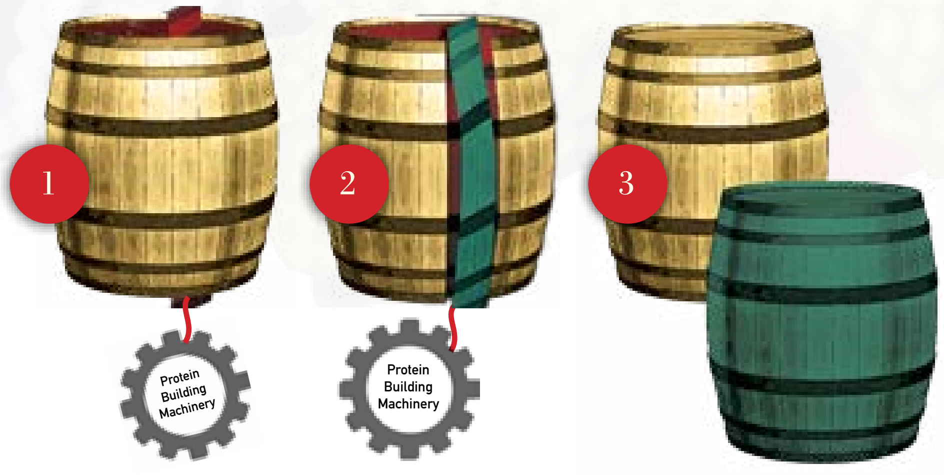
> “We’ve found that a lot of network users either have unrealistically high or unrealistically low expectations. With Web10G, we’re going to automate that process to let users know what’s reasonable and then help them get the performance they need.
> —Chris Rapier, network applications engineer

Roll Out the Beta Barrels



The Anton simulations allowed Gumbart and his colleagues to follow the motion of the beta barrel protein BamA over relatively long time scales. Above, Tan: exterior of the BamA "barrel"; red: interior; blue/gray: the molecules that make up the cell membrane.

Below, artist's conception of how the lateral opening in BamA may help insert other beta barrel proteins into the bacterial outer membrane.



Step 1: The strands, or "stays," of a beta-barrel protein can't be inserted into the membrane directly—it's not stable. The interior of the BamA channel stabilizes each stay as the protein building machinery makes it.

Step 2: As the Anton simulations showed, BamA parts to create an opening, allowing the other protein's strands to feed out into the membrane as they're made.

Step 3: Once the other protein's barrel is complete, it moves away, fully inserted in the membrane and ready to do one of a number of critical jobs for the bacterium's survival.

Anton Simulations Reveal How Dangerous Bacteria Install Critical Proteins

Why It's Important

In an era of diminishing antibiotic effectiveness, it's no wonder that bacteria, how they live—and what molecular components they can't live without—are an important focus for biomedical science.

This "beta barrel protein" inserts other beta barrel proteins into the outer bacterial membrane, including those that import nutrients or export toxins that kill host cells. The process is a promising target for antibacterial drugs.

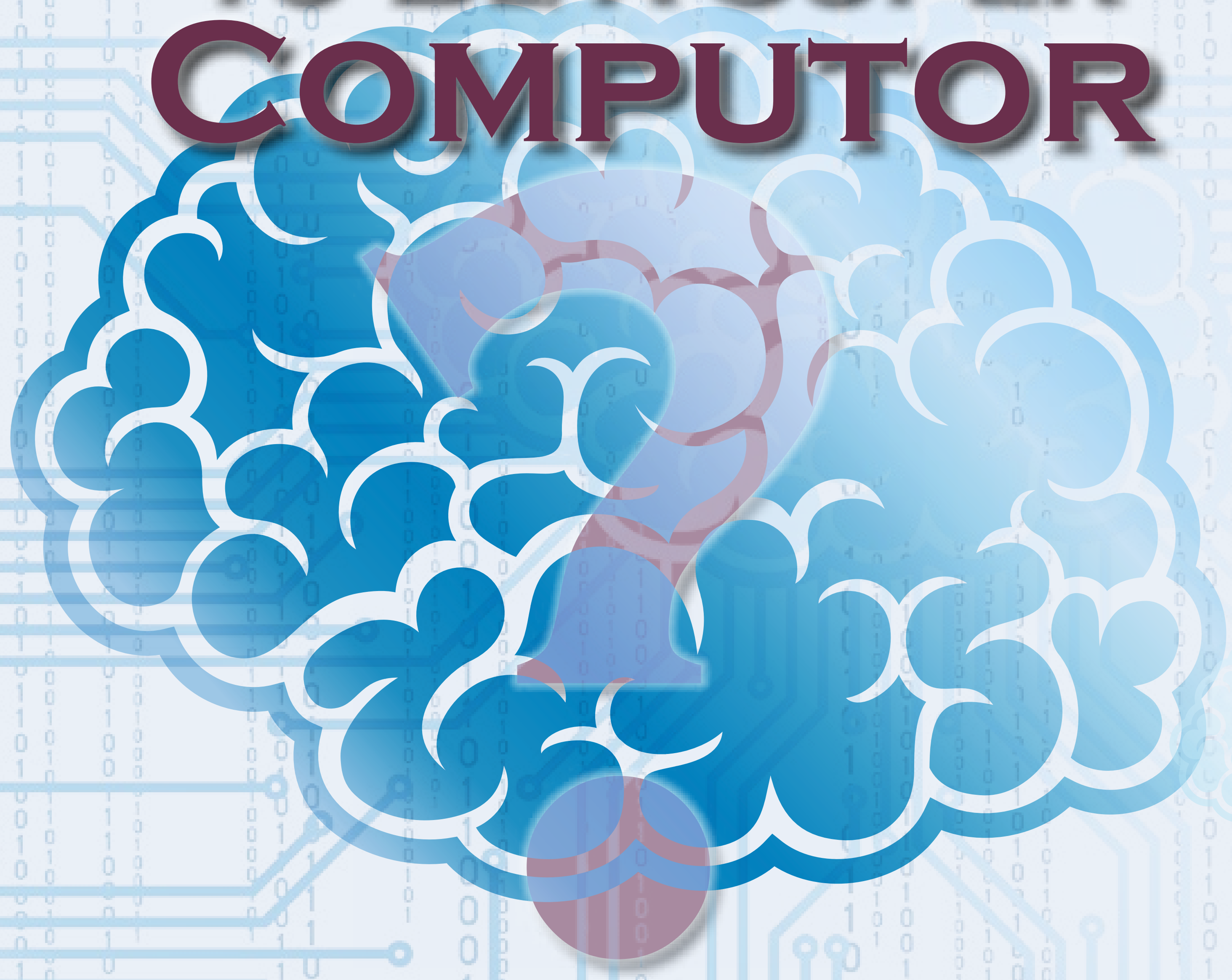
> *"When I saw the lateral opening in the simulation, I was surprised—even shocked. I never imagined beta barrels just opening spontaneously like that.*
> *—James C. Gumbart, Georgia Institute of Technology*

How Anton Helped

The researchers revealed BamA's side exit using molecular dynamics (MD) simulations that lasted from one- to two-millionths of a second. In the world of computational biochemistry, that's a very long time—supercomputers take months to perform simulations of the necessary length. On Anton, a special-purpose supercomputer designed to dramatically increase the speed of MD simulations, it can be accomplished in a day. The researchers reported their work in the journal *Nature*.

"Anton was critical for the work," Gumbart says. "If limited to conventional systems, I probably would have run about 50 to 100 nanoseconds"—a tenth or less the time scale. If he'd only looked at this scale, he says, he might have thought, "Well, I don't see anything, and that's what it is." Anton allowed him to push farther, to a remarkable result.

So You Want to Be a Super Computor



PSC Takes Lead in XSEDE Summer Research Experience Program

Young career seekers in the high-performance computing (HPC) field often face a familiar problem. You can't get the job without experience.

But another hitch confronts would-be "super computors" (pun intended). If you want to write software for PCs or smartphones, you probably know what that kind of programmer does. But what does an HPC engineer, researcher or educator do? How does a student find out if HPC is for him or her?

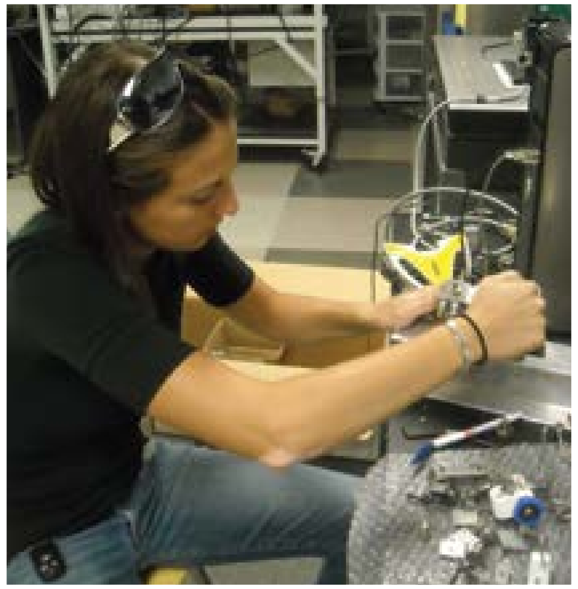
The NSF XSEDE Summer Research Experience Program exists to help students solve both problems, says PSC's Laura McGinnis, the program's coordinator. It also prepares and sustains a larger, more diverse pool of undergraduate and graduate students to be future HPC professionals. "Because supercomputing is a niche, we're providing a hands-on opportunity for students to experience HPC and be able to make an informed decision about their career track," McGinnis says. "We provide real-world experience in computational science, particularly for underrepresented students."

The Summer Research Experience Program includes training, internships, fellowships, mentoring and recognition activities. Participating students gain real-world research and development experience as well as encouragement and academic support in their pursuit of advanced degrees and digital services professional careers.

The program distinguishes itself from other internship programs by providing the participants the opportunity to expand their horizons with high-performance computing challenges in all research fields. Working with XSEDE researchers and staff, students gain relevant high-performance computing experience on real-world problems and the opportunity to make meaningful contributions to research, development and systems projects.

"It's important that the projects be real and not just have the students come in and optimize 'hello world' on a thousand processors," says McGinnis. The program also provides a small stipend and travel support for project orientation and attendance at the XSEDE14 conference in Atlanta, Ga.

PSC plays a central role in the Summer Research Experience Program, both by providing leadership and also by hosting many students in the program's summer study component. Here are a few of these students and their stories.



MARJORIE ADELE INGLE: FROM A LINE OF ROCKETEERS

Rockets are in Marjorie Ingle's blood. The University of Texas at El Paso (UTEP) second-year master's student literally learned rocketry on her grandfather's knee.

"He was a research and development engineer at White Sands Missile Range," she says. "He used to give me little models that I would put together. I 'volunteered' Barbie for the space program I don't know how many times," sending the doll roaring skyward on model rockets.

Ingle's XSEDE project centered on a phenomenon common to rocket engines as well as high-temperature nuclear reactor cooling systems: understanding how fluids such as rocket fuel or liquid helium coolant behave when they pass through small apertures.

Ingle worked with Pittsburgh Supercomputing Center's Anirban Jana on a computational fluid dynamics (CFD) model of jets of liquid helium flowing through a reactor cooling system. This system is under study partly because liquid helium coolant is more durable than the water used in older reactor designs and so doesn't need to be replaced as often, reducing the production of radioactive waste and making the reactor more ecologically friendly.

ANTHONY RUGGIERO: FINDING HIGHER POTENTIAL

For Anthony Ruggiero, a junior at Pittsburgh's Duquesne University, physics always seemed to be a gateway to higher things: find a higher potential, as it were.

"With physics, anything is possible," he says. "I wanted to do something with my life that people have never done before; I figured physics would allow me to do that."

Ironically, Ruggiero's XSEDE project literally focused on potential: the potential energy of an electron in what's known as the single-site Schrödinger equation.

In the equation named for him Schrödinger, one of quantum mechanics' founders, created the quantum equivalent of classical conservation of energy—an object's total energy is its potential energy plus its kinetic energy (the energy of its movement). In the strange quantum world, though, an electron doesn't have a location *per se*. Its location is more of a smeared-out cloud of possibility.

> *"Because supercomputing is a niche, we're providing a hands-on opportunity for students to experience HPC and be able to make an informed decision about their career track.*
> *—Laura McGinnis, PSC, XSEDE Student Research Experience Program coordinator*

Working with PSC's Yang Wang and Roberto Gomez, Ruggiero worked on speeding up calculations based on Schrödinger's equation via general-purpose graphics processing units—GPUs. Originally developed to help computers create smoother, more stable visual images, GPUs have proved extremely versatile even for calculations not related to images, such as those in Schrödinger's equation.

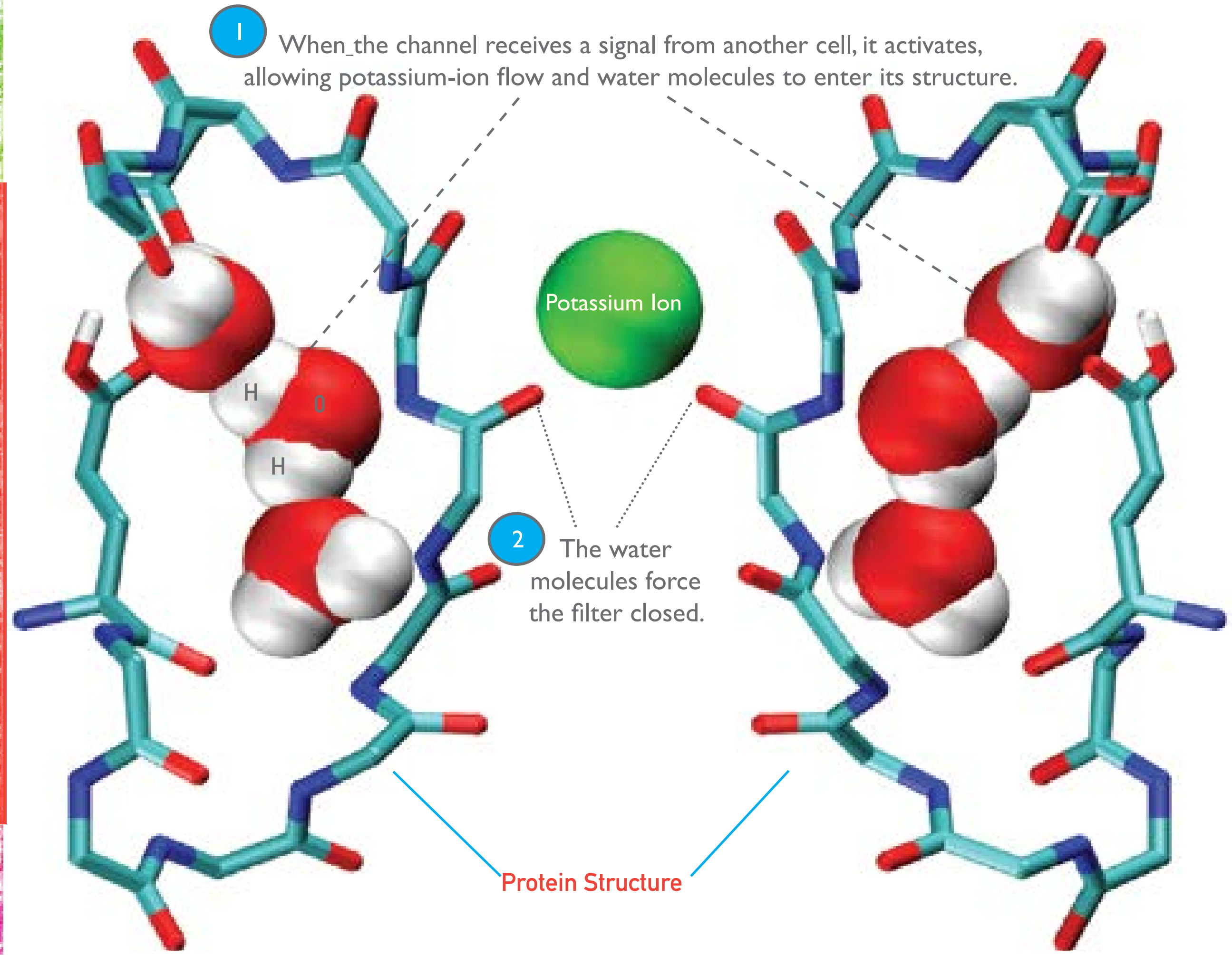
PAULA ROMERO BERMUDEZ: APPLYING A UNIQUE BACKGROUND TO PARALLEL COMPUTING

Paula Romero has had some unique educational experiences. When she was 11, the second-year University of Indianapolis undergraduate's family fled the political instability of Venezuela, where she was born, for the "old country"—her parents' native Spain. There she entered a teaching system very unlike that of the U.S.

"In Spain they focus on teaching theory," she explains. "They try to keep what is math on one side and what is physics on the other side… relating both fields is mostly your job. There is a lot of sitting down at a desk and studying for hours."

Conceptually, she says, such a parallelized education was great preparation for the mindset necessary for parallelizing code: pulling problems apart into chunks that can be attacked in parallel, speeding the calculation on computers with many parallel processors.

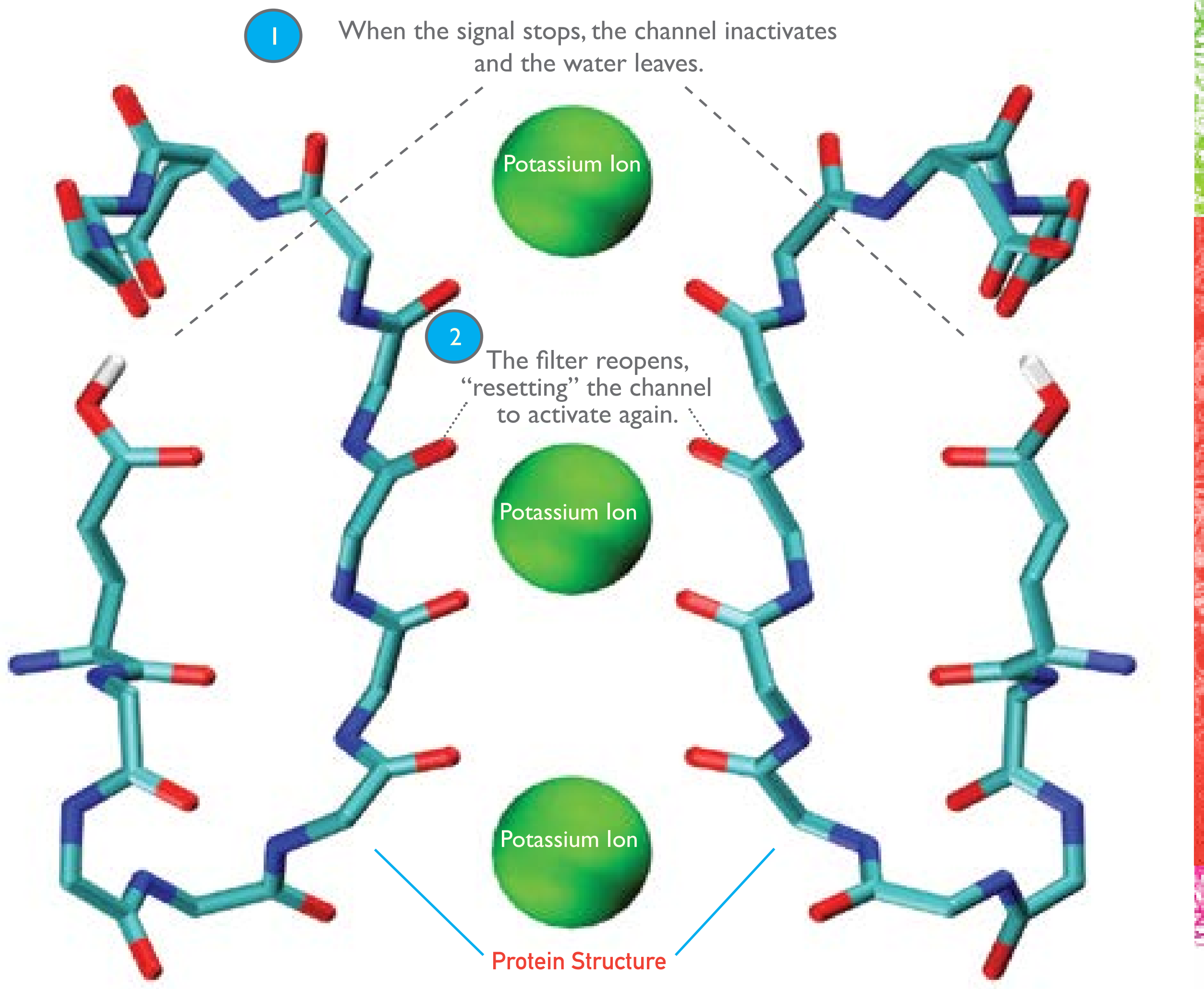
Under the guidance of PSC's Yang Wang and Roberto Gomez, she worked with Shawn Coleman, a PhD student at the University of Arkansas, to optimize an x-ray crystallography diffraction algorithm for use in Intel MICs—many-integrated core coprocessors, which speed highly parallel calculations in the Texas Advanced Computing Center's Stampede supercomputer.


1
When the channel receives a signal from another cell, it activates, allowing potassium-ion flow and water molecules to enter its structure.
H
O
H
Potassium Ion
2
The water molecules force the filter closed.
Protein Structure

The potassium channel activates nerve and other electrically active cells by allowing electrically charged potassium ions across the cell membrane. In its normal function, it then has to turn that flow of ions off – and then regenerate its original state so it can signal again. The researchers used Anton to see how the "channel" turns off and then back on.

Two Steps Forward, One Step Back

Anton shows how water leaving, re-entering potassium channel structure delays return to active state


1
When the signal stops, the channel inactivates and the water leaves.
Potassium Ion
2
The filter reopens, "resetting" the channel to activate again.
Potassium Ion
Potassium Ion
Protein Structure

WHY IT'S IMPORTANT

The potassium channel helps create electrical signals in nerve and muscle cells. This process goes awry in some irregular heartbeat conditions. To work properly, every nerve or heart cell needs potassium channels that can activate, inactivate and then reset themselves to respond to the next signal.

> *"Normally when you study molecular processes that take seconds to occur you imagine something of great complexity. Here we're talking about a process that's actually extremely simple.*
> *—Benoît Roux, University of Chicago*

In the journal *Nature*, Jared Ostmeyer, Benoît Roux and colleagues at the University of Chicago reported simulations on an Anton supercomputer, developed and provided by D. E. Shaw Research, hosted at PSC and funded by MMBioS, that revealed how the channel pinches off potassium movement.

HOW ANTON HELPED

In the cell, the potassium channel can take as long as 10 to 20 seconds to reset its potassium filter. No computer currently on Earth can carry out such a long molecular dynamics simulation. But the University of Chicago researchers leveraged Anton to push their simulations to 20 microseconds.

Even this relatively brief look was revealing, Roux says. "The system was stable for 20 microseconds in the pinched state," he says. That's a long time in molecular dynamics. "That was really shocking; we did not expect it." But that didn't mean the structure was static: The water molecules kept coming and going. Clearing the water molecules, and reopening the filter, was a "two steps forward, one step back" process, explaining the system's slow recovery.

NEWS IN BRIEF: PSC AND ITS PARTNERS

PSC RECEIVES FOUR NATIONAL AWARDS

In November, PSC received top national honors in four categories of the 2013 HPCwire Readers' and Editors' Choice Awards. HPCwire, the premier trade publication for the high-performance computing (HPC) community, announced the winners at the 2013 International Conference for High Performance Computing, Networking, Storage and Analysis (SC13), in Denver, Colorado.

PSC received:

- **Reader's Choice, Best use of HPC in Life Sciences**, for work on PSC's Blacklight supercomputer in overcoming limitations in complex DNA and RNA sequencing tasks, identifying expressed genes in nonhuman primates, petroleum-digesting soil microorganisms and bacterial enzymes that may help convert non-food crops into usable biofuels.

- **Reader's Choice, Best use of HPC in "Edge" HPC Application**, for the VecNet Cyberinfrastructure (CI). A collaboration between PSC's Public Health Group and Notre Dame's Center for Research Computing is building a computational system that will enable VecNet—a partnership of academic and industrial researchers, local public health officers and foundation and national decision makers—to test ideas for eradicating malaria before applying them in the real world.

Best Application of HPC in Financial Services
Best Use of HPC in "Edge HPC" Application
Best Applicat of Big Data in
Best Use of HPC Application in Life Sciences
HPC
wire
Editors' Choice Awards
Readers' Choice Awards
2013

- **Editors' Choice, Best Application of "Big Data" in HPC**, for PSC's newest supercomputing resource, Sherlock. Specially designed to solve what are known as graph problems, Sherlock is optimized for questions involving complex networks that can't be understood in isolated pieces. Topics range from cancer protein and gene interactions to performing smarter information retrieval in complex documents such as Wikipedia.

- **Editors' Choice, Best use of HPC in Financial Services**, for research that led to a change in trader reporting requirements to the New York Stock Exchange and NASDAQ. Work on Blacklight enabled investigators to prove that high-volume automated traders were exploiting reporting rules to make "invisible" trades that manipulated the markets.

PSC PUBLIC HEALTH EFFORTS MAKE TOP SUPERCOMPUTING DISCOVERY LIST

Two public health projects at PSC have also made HPCwire's list of "The Top Supercomputing-Led Discoveries of 2013." The HERMES project is analyzing public-health supply chains in lower-income countries to identify and repair under-appreciated choke points in vaccine supply efforts, for example. The VecNet Cyberinfrastructure project has created a prototype computational system to support a global malaria eradication effort (see below for more).

HPCwire named the two PSC projects among 30 supercomputing projects chosen from their new archives and which the publication believes are "set to change the world in 2014 and beyond."

PROTOTYPE CYBERINFRASTRUCTURE RELEASED FOR VECNET MALARIA ERADICATION PROJECT

In addition to winning an HPCwire award and being cited as one of the most significant supercomputing discoveries of the year (see above), VecNet CI has also completed its prototype of a computational system for university, industry, government and funding entities to test ideas for eradicating malaria.

"After our first year of development, we have a successful prototype framework of all user tools," says Nathan Stone, principal investigator of the infrastructure project at PSC. "Now, direct engagement with stakeholders via workshops, tutorials and online demonstrations will be important to refine these tools and get them into the hands of those who can best use them."

The final quarter of calendar 2013 saw the completion of the prototype CI framework, which consists of four major tools. These allow users to test existing and new malaria eradication methods, investigate malaria risk factors, plan detailed intervention campaigns and assess economic impact.

To unite these user tools, the group developed a common web interface with supporting access to a digital library (for archiving data sources and provenance), compute clusters (for running the disease forecasting models) and a data warehouse (for interactive access and analysis of calculated results). The web site and tools are now in use by almost 150 users worldwide from a variety of disciplines.

"Vecnet's work in 2014 will emphasize improvements to the malaria transmission models, and the expansion of calibrated input data to cover new geographic regions of interest to stakeholders," Stone says.

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 U. S. Department of Defense
The U. S. Department of Energy
D. E. Shaw Research
Cray Inc.
Silicon Graphics, Inc.
The Buhl Foundation
Bill and Melinda Gates Foundation

SENIOR WRITER: Ken Chiacchia
DESIGN & PRODUCTION: Shandra Williams
MANAGING EDITOR: Vivian Benton
PROJECT COORDINATION: Cheryl Begandy

COVER GRAPHIC: Simulation of the motion of the beta barrel protein BamA over relatively long time scales from research of James C. Gumbart of Georgia Institute of Technology and colleagues have been using an Anton supercomputer, developed and provided by D. E. Shaw Research, hosted at PSC and funded by MMBioS.

PRINTING: Heeter

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