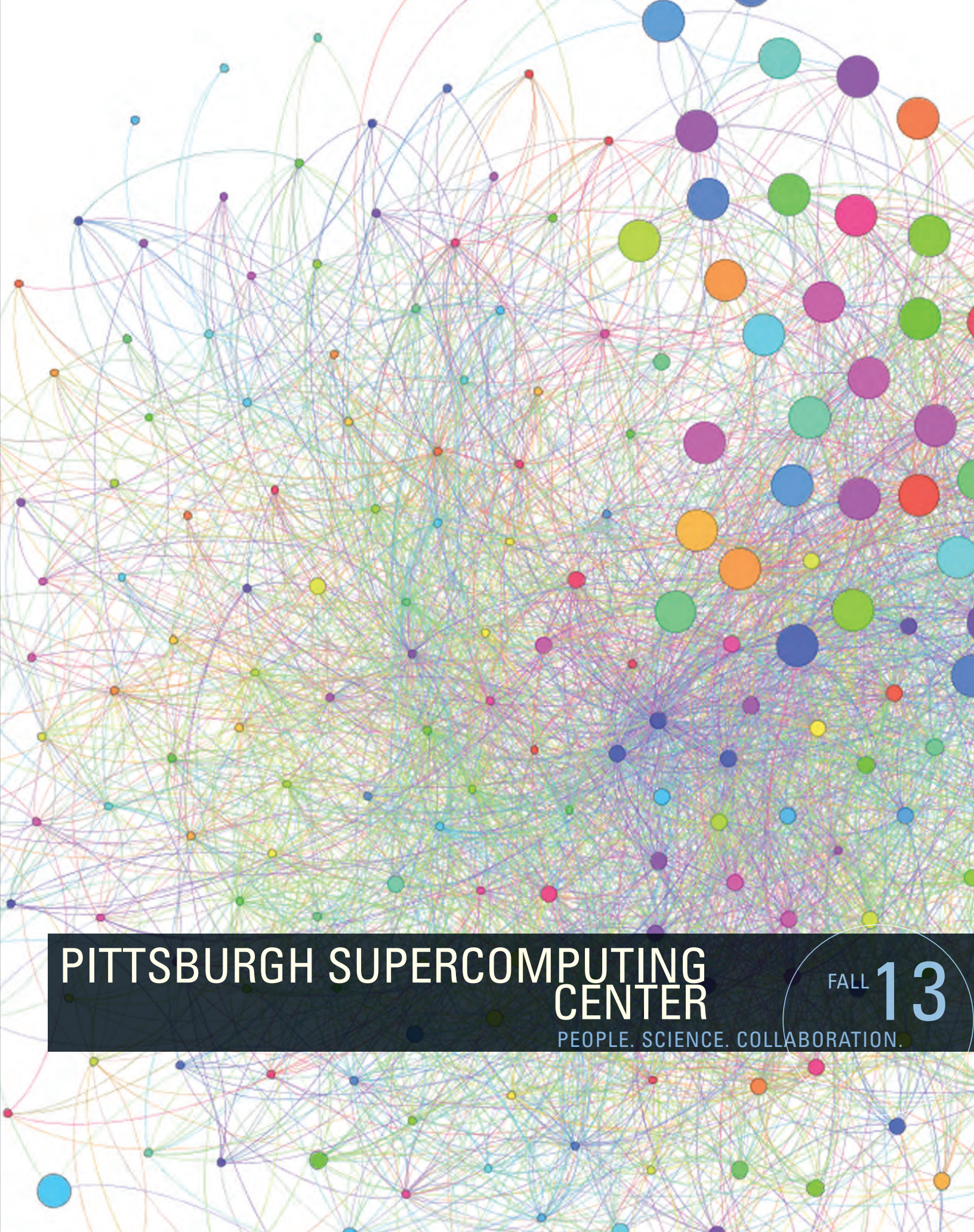
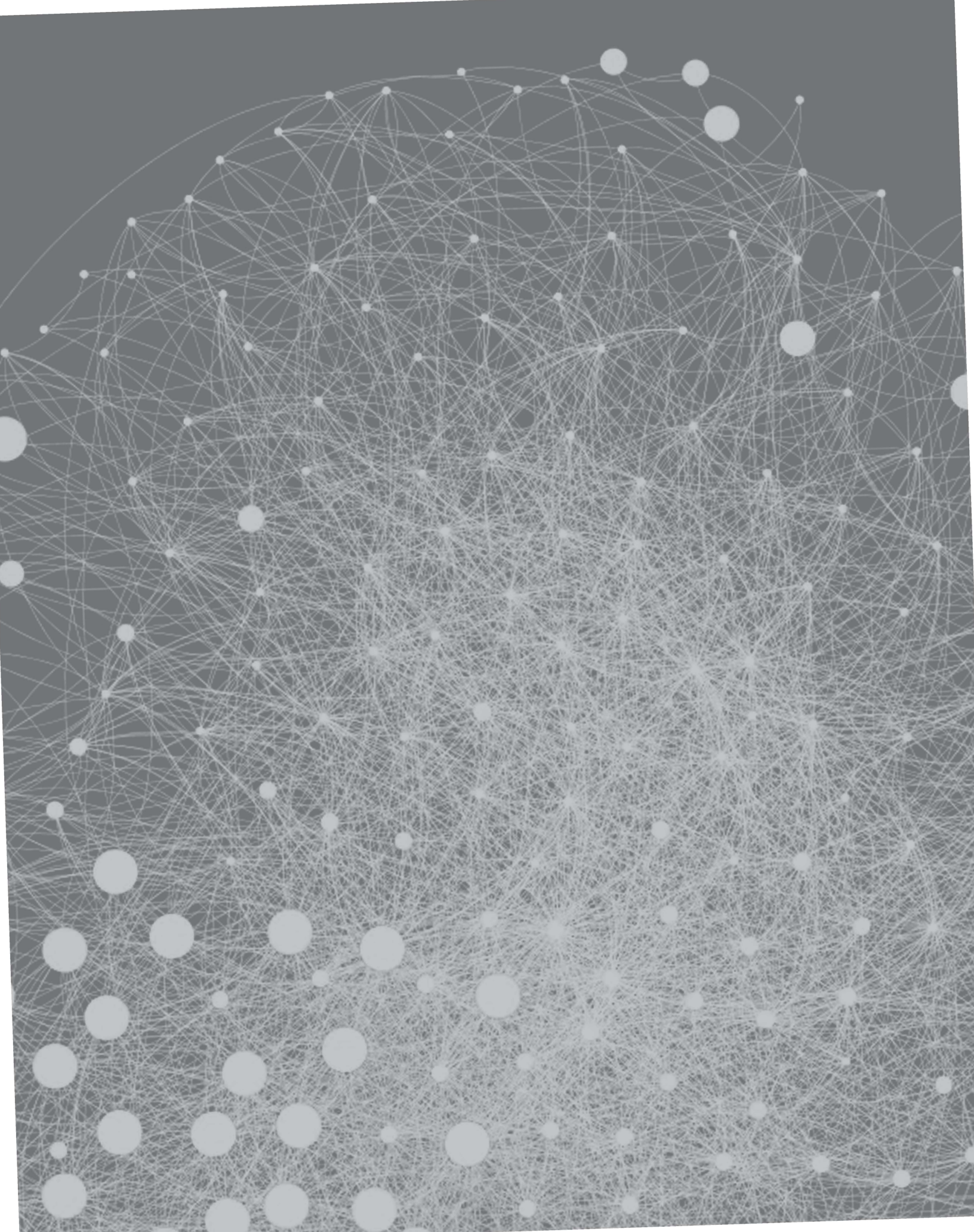


PITTSBURGH SUPERCOMPUTING CENTER

FALL 13

PEOPLE. SCIENCE. COLLABORATION.

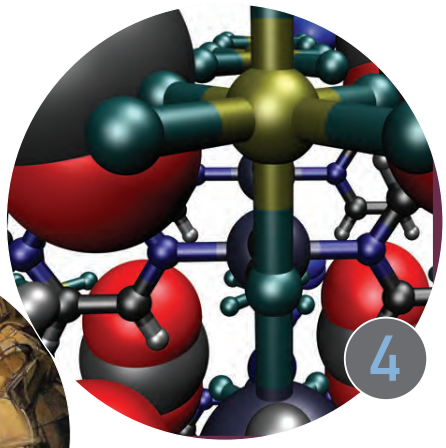


PSC.EDU

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

www.psc.edu
412.268.4960

CONTENTS



CHEMISTRY/ENERGY POLICY

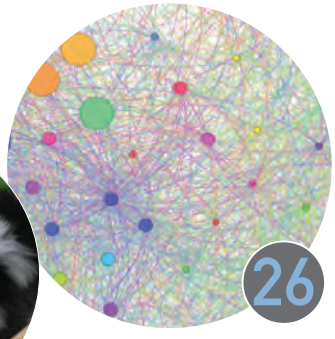
Building a Better Carbon Trap

Brian Space and colleagues at Florida State University and elsewhere use **Blacklight** to help design *materials for capturing carbon dioxide* from factory, vehicle exhaust.

BIOCHEMISTRY/NEUROSCIENCE

A Movie is Worth a Million Pictures

Anton animations of *key nerve-cell* protein's structure give Elia Zomot and Ivet Bahar of the University of Pittsburgh insight not possible with previous static images.



PRIMATOLOGY/MOLECULAR BIOLOGY

Concentration

Non-Human Primate Reference Transcriptome Resource at Weill Cornell Medical College and the University of Washington uses **Blacklight**, **Data Supercell** to *identify pivotal genes in primates.*

LINGUISTIC FORENSICS/DATA SEARCH

Needles in a Needlestack

PSC's **Sherlock** Supercharges Next-Generation Search Tool.

FROM THE **DIRECTORS**

We're pleased once again to highlight the world-class work being done at PSC—in scientific research, development of research infrastructure and education—by collaborating researchers, staff members and students.

In the following pages, we focus on the value our efforts bring to pressing national issues such as energy policy (p. 4), biomedical advances (p. 18) and even the stability of the stock markets (p. 15). One exciting facet of our accomplishments over the last year has been in the public health arena (p. 16). Importantly, PSC's advances have contributed to the larger community at multiple scales, offering benefits at the national, state and local levels.

PSC's ability to attract scientific research dollars into the state is exemplified by the five-year, $9.3-million National Institutes of Health grant that funded the multi-institutional Multiscale Modeling of Biological Systems (MMBioS) program, which you can read about on p. 11. Accompanying it is a feature article on how our Anton supercomputer, from D.E.Shaw Research, helped correct biologists' understanding of a key protein in brain function—a major early focus for MMBioS.

PSC's other systems, which are a part of the national cyberinfrastructure, support a wide range of research and education efforts with their unique capabilities. *Blacklight*, from SGI, provides the largest nationally-available cache-coherent memory vital for biomedical and other projects. *Sherlock*, a Urika™ system from YarcData, a subsidiary of Cray, is uniquely purpose-built for data and graph analytic information analysis. Our novel *DataSupercell* data storage system, designed and built at PSC to provide the capacity and performance required for data-intensive work, has replaced our tape-based archive. These systems are described in detail on our webpages (www.psc.edu).

On p. 12, we chart how PSC played a pivotal role in migrating the communications network of the National Science Foundation's XSEDE collaboration of computing sites to Internet2. This upgrade has begun enabling ultra-broadband connections at the XSEDE sites. The PSC expertise that made this work possible also benefits the local and regional universities and research institutions served by our Three Rivers Optical Exchange (3ROX).

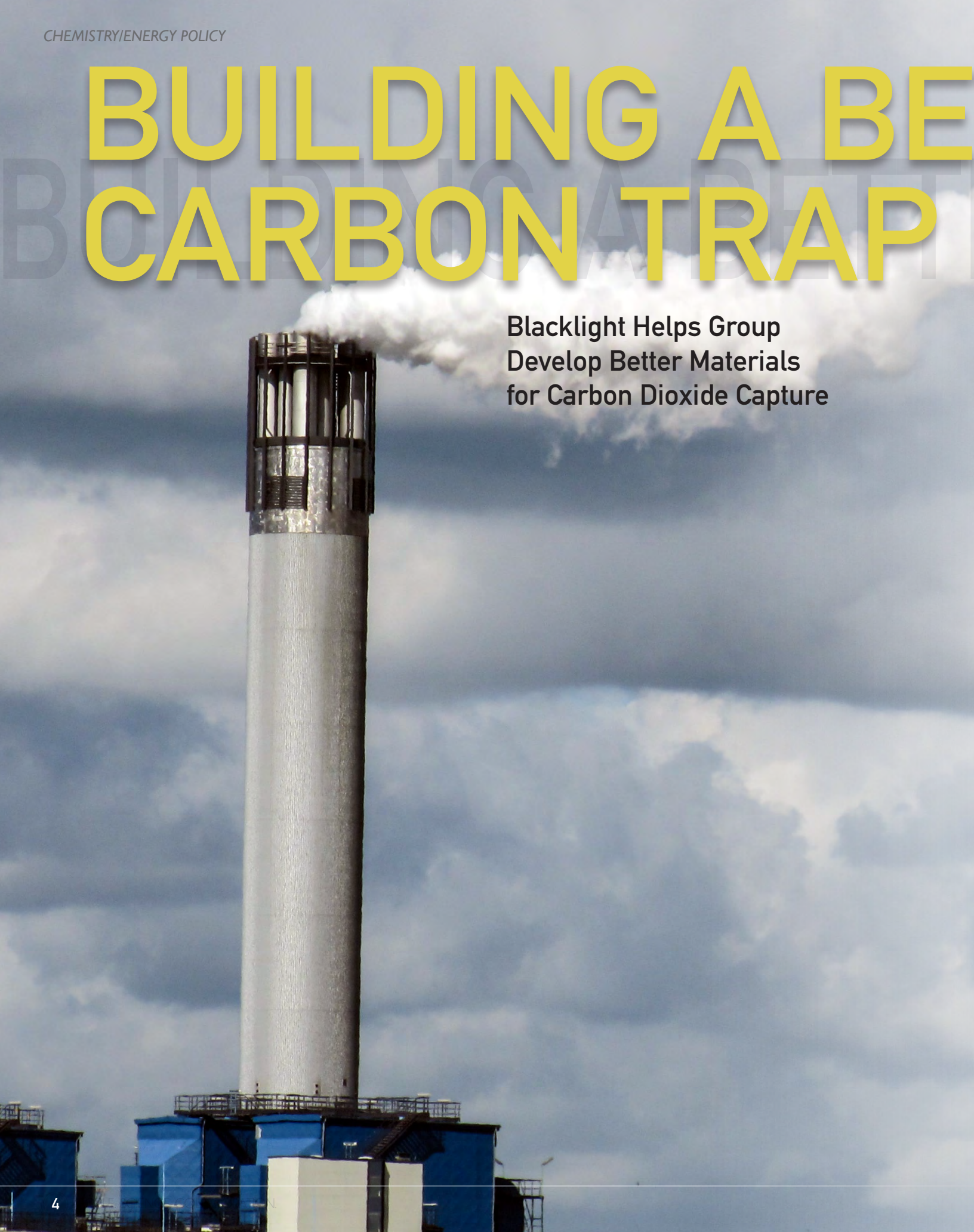
Our educational programs help develop a sophisticated, technically capable workforce that benefits the city, region, and state. An article on our Minority Access to Research Careers program, on p. 22, features one story—among several—about how our educational efforts led a talented young man from a City of Pittsburgh senior high school project at PSC to a summer job here, and now to higher education in computer science. See p. 14 for additional state and regional impact.

These are just a few of the PSC success stories described in the following pages—see the Table of Contents for others. Stay tuned for more six months from now: you'll notice from our cover that we've increased the frequency of PSC's previously annual report. *Pittsburgh Supercomputing Center* magazine will now appear twice per year, helping us enhance our communications and broaden our means for demonstrating our value to the government agencies, organizations and companies that fund and support our work—and, not incidentally, to those who contribute to our nonprofit mission at psc.edu/donate.

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

BUILDING A BE
CARBON TRAP

Blacklight Helps Group Develop Better Materials for Carbon Dioxide Capture



TTER ER CARBON TRAP

In May 2013, a U.S. government lab in Hawaii measured a carbon dioxide concentration of over 400 parts per million. It was the first time in the lab's 55 continuous years of operation that the gas reading was that high. In fact, the last time it topped 400 ppm, *Australopithecus afarensis* walked a world that was more than 2 million years away from seeing its first modern human.

Human generation of carbon dioxide and other greenhouse gases is altering the global climate. But today's world is very dependent on carbon-dioxide-generating fossil fuels. How do we make our economy "carbon neutral" while still *having* an economy?

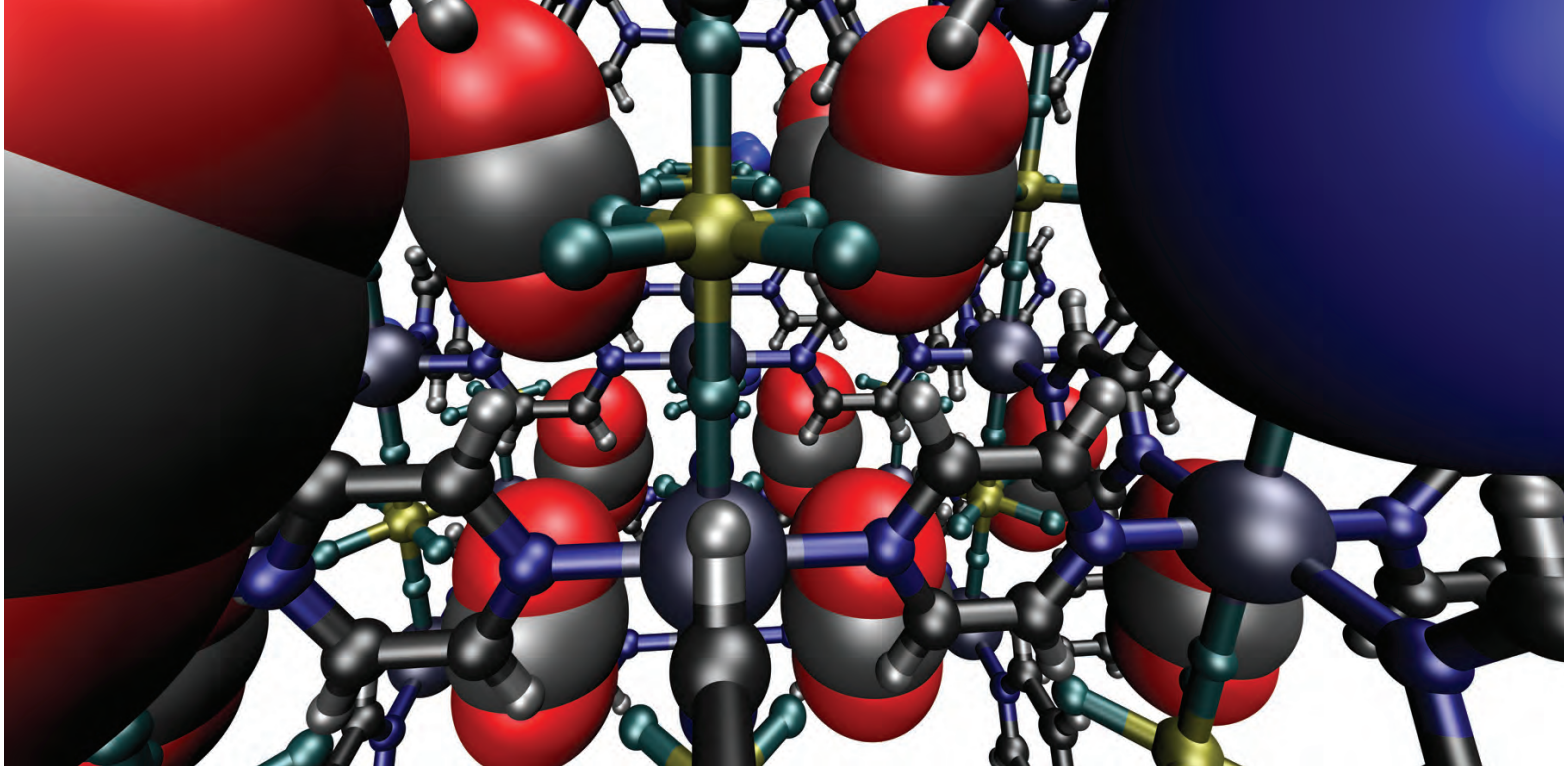
No one can say yet what technologies we'll need to solve the climate change dilemma. But capturing carbon dioxide from smokestacks and other waste streams is attractive. This is because it would allow us to continue using fossil fuels at least for a while.

In a February 2013 online *Nature* paper, a team led by Patrick Nugent of the University of South Florida's Department of Chemistry and Youssef Belmabkhout of the King Abdullah University of Science and Technology's Advanced Membranes & Porous Materials Center has reported a new material that may fit the bill for carbon capture. They have developed a series of metal organic frameworks (MOFs) that may be able to "filter" carbon dioxide out of smokestacks and exhaust pipes.

"The new materials show very selective carbon dioxide capture," says co-author Brian Space, professor of chemistry at the University of South Florida. "Indeed, carbon dioxide fits [into the MOFs] like a glove; nothing else fits as well."

Space's team used supercomputers to simulate how passing through an MOF affects exhaust gases. These simulations allowed them to explain the behavior of MOFs and give their lab-researcher colleagues the clues they needed to create better MOF "carbon scrubbers."

Pittsburgh Supercomputing Center's Blacklight, an SGI UV shared-memory system, Space says, was crucial for an initial, detailed series of simulations. These showed how previous simulations had missed something important. The new simulations also pointed to the way forward.



Carbon dioxide molecules (gray and red) trapped in the MOF matrix.

DETAILS COUNT

Space's group attacked the simulation in steps. The many gases in an exhaust flow and the components of the MOFs could be simulated accurately only after the investigators understood the details of their interactions. And previous models were missing something. They didn't make good predictions of which MOFs would attract which gases.

Space and his team produced exact simulations of individual gas molecules and tested how they interacted in pairs and triplets.

Moving from simulating pairs of molecules—which earlier investigators had done—to triplets really increased the demand on computer resources, Space says. "It's difficult to calculate." But the researchers discovered that adding the third molecule made a big difference in the computers' predictions. And it made them *right*.

Inside the MOFs, Space and his team discovered, electrical charges were distorting the gas molecules, attracting them much like a magnet attracts iron filings. Previous simulations hadn't included such "polarization," partly because two-molecule models had suggested it wouldn't be important. But when the simulation team added the third molecule, they found that polarization's effect was much stronger than expected.

Only Blacklight, which is the largest "shared-memory" computer in the world, had the memory capacity to perform the triplet calculations. (*See Technical Note, p. 7.*)

"These energy models were not even possible a few years ago," Space says. "We can now explain the experiments at an unprecedented level of accuracy."

GIVE-AND-TAKE TOWARD A BETTER CARBON TRAP

Space's group worked with the laboratory researchers in a give-and-take way.

"We start out by asking, 'Why does this work this way?'" he says. "Then we ask, 'How do you improve this—how would you change the way it works?'"

With the improved simulations, Space's team could better explain the properties of known MOFs. This knowledge allowed them to experiment—in the simulations—with different MOF structures to see if they would bind carbon dioxide better. The lab researchers created these MOFs and measured their real properties. Space and his team then plugged differences between the predicted and measured properties into their simulations. In the end, the back-and-forth led to even better predictions.

The work identified a family of MOF structures containing an electrically charged silicon-fluoride compound that attracts carbon dioxide much more strongly than other gases. They even work in the presence of water vapor, which prevented efficient carbon capture in earlier materials. This discovery was important enough to merit an article in *Nature*, one of the world's most pre-eminent research journals.

"It's an exciting time in the field," he adds. "The level of accuracy, if it's done carefully, is really predictive now."

/ Technical Note

Harnessing Memory to Make the Trap Work

Brian Space and his collaborators needed a number of supercomputing resources to move from modeling the behavior of pairs and triplets of gas molecules to bulk properties of gaseous mixtures and their interactions with candidate metal organic framework (MOF) matrices. Perhaps not surprisingly, each of these modeling steps required a different type of computational resource.

"For the different pieces of the puzzle we needed different machines," Space explains. "After our group develops an in-depth model of the forces between various guest molecules and their host materials, we perform molecular simulations of the sorption-mediated processes on highly parallel [supercomputing] resources." His group used a number of machines in the National Science Foundation's XSEDE network, including TACC's Ranger, SDSC's Trestles, and GPU clusters, such as Georgia Tech's Keeneland for the latter calculations.

But the initial simulation of small numbers of molecules, requiring exquisitely detailed quantum mechanical modeling, is so memory intensive that only one XSEDE resource—and one program—made sense: PSC's Blacklight, running Molpro.

"Their simulations require a lot of memory, so they're ideal for Blacklight's shared-memory architecture," says Marcela Madrid, senior scientific specialist at PSC. "Some of their runs took about a terabyte of memory; one of the problems was to tune the job script so that you would really get the memory you want and not get errors due to insufficient memory."

The ability to run Molpro on Blacklight was also important, Space adds. "Other programs would have taken a year or two to perform the calculations; with Molpro it only took hours."

"One issue was that Molpro is very I/O intensive," Madrid says. "We had to efficiently use the file system in order to accommodate this."

Madrid and Rick Costa, who is a staff computational science consultant at PSC, were instrumental in making it all work, Space says. "Without their help, we couldn't have done it."

"Science is really a cooperative effort," he adds. "Many years of people's work went into writing the codes to do these calculations at all, let alone efficiently. No one group can be an island; you have to depend on others."

A Movie is Worth a Million Pictures

Animation Corrects Understanding of Key Nerve Cell Protein

Duchamp's iconic *Nude Descending a Staircase, No. 2*, invites us to think about motion. By presenting a mundane action in many sequential frames, presented all at once, it fundamentally alters the way we'd think about the action hinted in a static image.

It's much the same with the mechanisms of complex molecules inside living cells. In the case of the aspartate and glutamate neurotransmitter transporters of the brain's nerve cells, static X-ray pictures had offered a fairly compelling argument for how the transporters work. Trouble is, the pictures were misleading.

Using simulations run at PSC on a special-purpose supercomputer called Anton, Elia Zomot and Ivet Bahar of the University of Pittsburgh School of Medicine have created "movies" of the transporter that show it simply can't move the way that the static images had hinted. These data provide important new insight into the mechanism of an important actor in removing neurotransmitters from the space between nerve cells, a phenomenon with relevance to medical conditions involving the progressive death of nerve cells, like stroke or Alzheimer's disease. The researchers reported their results in the *Journal of Biological Chemistry* in November 2012.

A VERY INTERESTING MACHINE

The space between a nerve cell and another nerve cell to which it communicates is called the "synapse." When a nerve cell fires, it floods the synapse with chemical messengers called neurotransmitters. This in turn causes the second cell to fire. The transmitter molecules remaining in the synapse afterward, though, pose a challenge to the nerve cells.

"It's important to clear excess neurotransmitter from the synapse," says Bahar, John K. Vries Chair in Computational & Systems Biology at University of Pittsburgh. "When you have such an excess ... you can develop neurodegenerative diseases" like Alzheimer's, Huntington's, epilepsy or the nerve-cell death following stroke or brain trauma.

The glutamate and aspartate transporters are proteins that form channels across the nerve cell's membrane. These transporters pump either glutamate or aspartate out of the synapse and into the cell, using the flow of sodium ions into the cell as a power source, says Zomot, first author of the journal article and a research associate in Bahar's laboratory.

The protein has two faces, which pivot open and closed, a little like the opposite ends of a clothespin. It starts with its outside-facing end open. When both a neurotransmitter molecule and two sodium ions attach to binding sites in the outward facing part of the channel at the center of the transporter—about where the spring is in a clothespin—the outward face swings closed. This motion opens the transporter's inward face, releasing both transmitter and sodium ions into the cell.

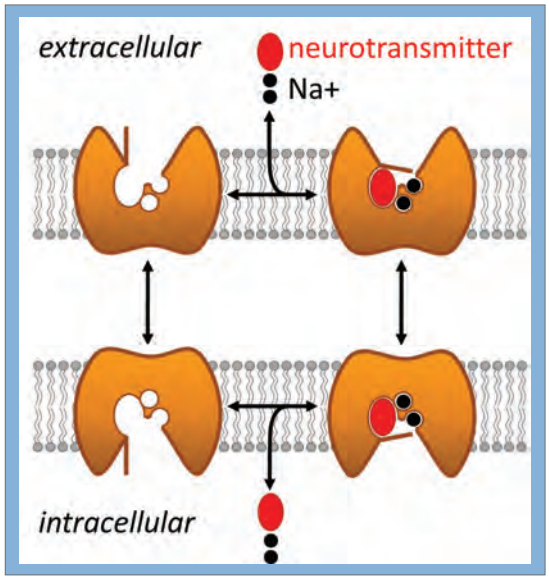
"It's a very interesting molecular machine," Bahar says of the transporter molecule. "It opens up, lets the transmitter in, then closes down, changing conformation and becoming inward facing so it can release the transmitter without leakage."

DEVIL IN THE DETAILS

The problem, though, was in the details. In the static X-ray images, a loop in the transporter protein's structure, called HP1, appears initially to block the channel near the central pivot point. If that were true, HP1 would work by moving aside and letting the transmitter and sodium ions into the cell when the protein pivots.

"People made inferences that this region was probably functioning as a gate," Bahar says. "But this was based only on static images."

Zomot and Bahar used the Anton supercomputer, which was designed and constructed by D. E. Shaw Research, to simulate the motions of the transporter as it pivoted. D. E. Shaw Research provided the Anton computer at PSC without cost



The transport cycle that the Glu/Asp transporter protein uses to empty the synapse of excess neurotransmitter. The protein starts with its transmitter and sodium binding sites facing outward ("extracellular," upper left). After the neurotransmitter (red oval) and sodium ions (Na+ black circles) attach to the transporter, the external gate closes, blocking them from escaping (upper right). Then the transporter reorients to face the inside of the cell ("intracellular," lower right). Finally, the internal gate opens, releasing neurotransmitter and sodium into the cell (lower left). The now empty transporter reorients to face the external space again, for a new transport cycle to start.

for non-commercial research use by the national biomedical community. Anton is hosted by PSC's National Resource for Biomedical Supercomputing, with support for operational costs provided by a grant to the National Center for Multiscale Modeling of Biological Systems, a National Institutes of Health-funded collaboration between PSC, University of Pittsburgh, Carnegie Mellon University and the Salk Institute. (*See Bridging Scales, p. 11.*)

Anton was very different from other systems Zomot has used. But learning its ways offered serious advantages, he says. "Once you become familiar with it, you find out that it's really worth the effort of learning the system. You can accomplish much, much more than you could with the other supercomputing clusters usually available."

Bahar agrees. "In a few hours, the machine generates what another would in months."

WORTH A MILLION PICTURES

Anton created a virtual model of a specific member of the glutamate/aspartate transporter family, an aspartate transporter called GltPh. Using ball-like atoms connected by spring-like molecular bonds, the computer calculated how all the different parts of the transporter moved around as they experienced atomic movements and vibrations.

The simulations delivered a bit of a surprise: HP1, which had been a favorite bet to be the gate based on the X-ray work, was not in the way (above). Instead, another loop, HP2, did seem to function as the gate. This is an important discovery potentially saving countless hours of drug development revolving around the wrong molecular target.

"Now we know really how it functions," Bahar says. "If you want to understand function, you need to see how it moves... If a picture is worth a thousand words, a movie is worth millions of pictures."

The researchers are getting ready to study another neurotransmitter transporter—one similar to those that move the brain neurotransmitters dopamine, serotonin and others out of the synapse.

"We have another allocation on Anton," Zomot says, and studies of the leucine transporter, which is more distantly related to the glutamate/aspartate transporters, could reveal both common mechanisms and useful differences for drug developers wanting to affect one but not the other. With Anton, "finally we have the option to do this type of study in a reasonable amount of time."

BRIDGING SCALES

MMBioS to Fill Gaps Between Biological Research Scales

While Elia Zomot's and Ivet Bahar's work with Anton has generated a surprising, corrective look at how a key nerve-cell protein works, increasingly computational biologists are concerned about the gaps in their vision.

Historically, computational biologists have worked at three different levels: the atomic level that determines how individual biomolecules work; the subcellular level that governs phenomena like metabolism and neurotransmission; and the whole-cell and tissue level that defines a functioning organism. The computational methods used to examine these phenomena leave big gaps between them, and it is in exactly those blind spots that some of the most important questions in the field now lie.

"If you look at how we approach biological problems using computation, it's very compartmentalized," says Markus Dittrich, director of the National Resource for Biomedical Supercomputing at PSC. "All these approaches help you understand problems at certain size and time scales, but one of the grand challenges in the field is to have a comprehensive treatment that spans all these approaches."

New methods and technologies are needed to bridge the information gained at each level of investigation—a rationale that underlies the new National Center for Multiscale Modeling of Biological Systems (MMBioS), a collaboration between the University of Pittsburgh, Carnegie Mellon University, the Salk Institute for Biological Studies, PSC and colleagues elsewhere. The new center started its work in September 2012 by means of a five-year, $9.3-million grant from the National Institutes of Health.

"MMBioS is multi-scale," says Bahar, who is John K. Vries Chair of Computational & Systems Biology at the University of Pittsburgh School of Medicine and the project's principal investigator. "It combines methods at the molecular, cellular and even tissue levels" to truly integrate events at all those scales.

MMBioS has three focus areas:

- molecular dynamics and supramolecular signaling and machinery, led by Bahar
- signaling networks at the cellular and sub-cellular levels, co-led by James Faeder, associate professor of Computational & Systems Biology at University of Pittsburgh and Terrence Sejnowksi, director of the Computational Neurobiology Laboratory at the Salk Institute
- cellular and tissue level image analysis and modeling, led by Robert Murphy, director of the Lane Center for Computational Biology at Carnegie Mellon

"Part of the vision behind MMBioS was that there would be strong connections between the projects," says Murphy. "I'm very excited about working together with such a group of very talented people."

For example, Murphy's earlier work included using machine learning to generate models of whole cells from microscopic images. One of the goals of the new center is to create a bridge between the efforts of the modeling and image analysis groups. A side goal will involve leaders in both communities to create standards for linking cell structural models with biochemical modeling tools.

"The Center will not only bridge the gap between computational biologists working at different scales," Bahar says. "It will also help establish a common language, and promote collaboration, between experimental and computational scientists currently tackling the same challenging problems with little help from each other."

MMBioS will begin with a concerted effort to develop technologies that enable neuroscientists to synthesize what they know about the nervous system. The eventual goal will be to combine knowledge about the chemistry of neurotransmitter transporters like the ones Zomot and Bahar are studying, the anatomy of and communication between brain cells and resulting phenomena such as intelligence and decision-making.

OPENING THE FLOODGATES

Argonne, PSC Staff Shepherd Internet2 Migration, Give XSEDE Network Bandwidth Needed for Big Data Era

In 2006, a senior U.S. Senator made the mistake of referring to the Internet as "a series of tubes." He instantly became the brunt of jokes about a guy who grew up in a time when people communicated via post, in cursive script, trying to make sense of an email world. But to be fair, it isn't such a bad metaphor.

Information—data—is as critical to our economy and society as fresh drinking water is to our homes. Like the plumbing running through our houses, the Internet transports data through "pipes" that are limited both by their size and by the capacity their "faucets" can deliver.

Thanks to personnel at Argonne National Laboratory and PSC—chiefly Linda Winkler, senior network engineer, Argonne; Joseph Lappa, principal network design engineer, PSC and Kathy Benninger (pictured), network performance engineer, PSC—the XSEDE network now has the "pipe capacity" it will need to keep up with the pace of the Big Data era. XSEDE, the National Science Foundation's U.S.-wide network of supercomputing sites, which includes Argonne and PSC, has accomplished this by migrating its data network to Internet2, a vastly higher-capacity system than the previous carrier. XSEDE's improved network will enable sites to achieve connection rates of up to 100 Gigabit per second (100 GE)—10 times faster than currently possible. The architecture of the new system will also allow a number of upgrades that will help the transfer of data through the system.

As part of the Internet2 migration, Lappa has taken on new responsibilities for the XSEDE network. Newly appointed as XSEDE's operations networking manager, he will be XSEDE's main contact with Internet2. In this role, he and his team will monitor the performance of the new network, oversee details of transitioning sites to 100 GE, assist with campus bridging, and help Internet2's programmers and service representatives optimize and tailor the network to XSEDE and its users' needs.

THE APPROACHING BOTTLENECK

Users at XSEDE sites employ some of the largest, fastest computers in the world to generate vast volumes of data. Moving those data between researchers, the supercomputers and storage sites is no small mission. To accomplish that job, XSEDE's predecessor, TeraGrid, originally built what was then one of the highest-capacity, most reliable networks in the world.

"Advanced networking is critical ... to support the researchers and educators who are making innovative use of our ... resources," says John Towns, XSEDE project director, noting that XSEDE supplies about 8,000 users with 17 supercomputers, data storage and management tools and networking resources.

In the Information Age, though, technology ages quickly. As the XSEDE network and its demands grew, it began to approach the limits of its infrastructure: in particular, a potential bottleneck between XSEDE sites in Denver and Chicago loomed large.

"As far as the technical reasons for migrating to Internet2, it was the 'speeds and feeds' problem," Lappa says. A factory, for example, can perform an operation on a product quickly (speed). But if it can't then move the next product up the line (feed) fast enough, that speed is wasted. Similarly, the blinding speed of XSEDE's computing machines was in danger of being made far less relevant by the approaching difficulty of getting data into and out of them.

UNCLOGGING THE PIPES

Internet2's 100 GE backbone proved to be the solution to the problem, Benninger says. "With 100 GE, there is a clearer path to allow us to operate."

While not all the sites will initially have 100 GE connections to the new backbone, she adds, the system will have room to grow to meet the next three years' needs. Currently, Indiana University and Purdue University share a 100 GE connection. A number of other sites plan to upgrade over the next several years.

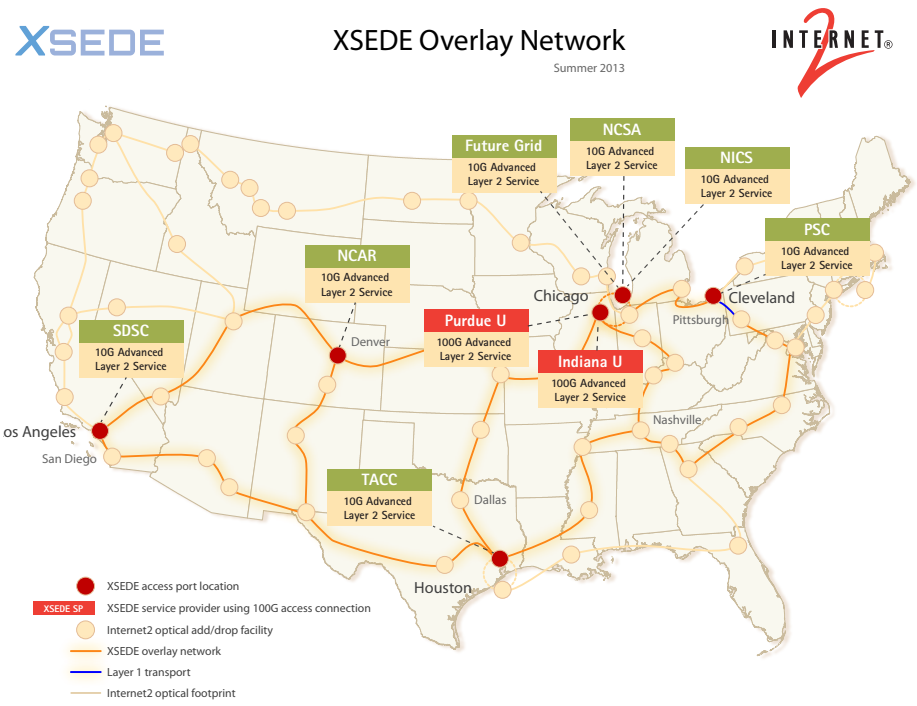
In addition to supplying the leadership for the migration process, PSC also served as one of the first sites on the new network, testing out and helping Internet2 improve and customize the system to serve XSEDE's needs.

Internet2's architecture offers a big plus in terms of managing data flow with what's known as "dynamic provisioning capability." If a particular network path between two sites is congested with large data flows, a network engineer can establish a new virtual local area network (VLAN) to route additional data transfers over an alternate path.

FUTURE UPGRADES

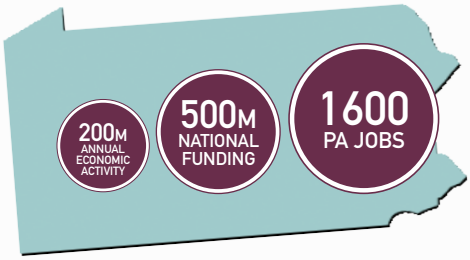
In addition to optimizing the network and helping sites connect with the backbone or upgrade to 100 GE, Benninger and Lappa will support efforts by a number of PSC and XSEDE staff to add new functions that take advantage of the higher bandwidth.

- The XSEDE-wide File System (XWFS) will allow the increasingly large files required by researchers to be moved rapidly between XSEDE sites.
- Web10G, developed by Chris Rapier, PSC network programmer, Andrew K. Adams, PSC network engineer, and John Estabrook, network programmer at the National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, will monitor data flowing from servers to the network to help pinpoint sources of slowdown even as they happen.
- VLAN (virtual local area network) provisioning will allow any two XSEDE sites to set up a "virtual network" between the two sites that performs as if it were a direct, hard-wire data connection, avoiding the need to set up potentially complex routing through the network.



NEWS IN BRIEF: PSC AND ITS PARTNERS

2014 PENNSYLVANIA STATE BUDGET INCLUDES $500,000 FOR PSC



The Commonwealth of Pennsylvania budget signed by Gov. Tom Corbett on June 30 included a $500,000 line item for PSC.

"This is very good news for PSC and for the Commonwealth," says Ralph Roskies, Scientific Director for PSC, adding that the state's return on its past investments in PSC has been excellent. "We're grateful to the members of the General Assembly, and especially the Allegheny County delegation."

Since PSC's inception, the center has brought over $500 million in outside funds into Pennsylvania, representing a 14:1 return on the state investment. PSC is also responsible for generating 1,600 jobs and over $200 million in annual economic activity.

Besides the direct monetary impact of PSC activities, PSC also enables scientists and engineers throughout the State to carry out research that would otherwise be infeasible. As a result they are more competitive for their own federal grants which adds to the Commonwealth's economic activity.

PSC's outreach and training efforts from high school and up advance STEM competencies within the State, leading to a workforce more prepared to exploit 21st century technology. Its corporate program enables Pennsylvania corporations to become more competitive.

And the results of PSC research contribute directly well-being within the Commonwealth as when PSC's Public Health group worked directly with national and local health departments to develop strategies for mitigating the spread of the H1N1 virus.

"In our fight for federal awards, we're competing with some of the best high performance computing centers in the world, many of which enjoy significant state funding," Roskies says. "The state line item will help us compete on a more even footing."

PSC, NUMASCALE A/S TO COLLABORATE ON IMPROVED MEMORY SYSTEMS FOR RESEARCH

PSC and Numascale A/S, whose products support the construction of low-cost, scalable-server computer systems, have launched a collaborative project investigating the applicability of Numascale systems to the many research projects requiring more directly-addressable memory than is readily available on single, commodity, multi-socket, large memory servers.

"Rapid advancement in many scientific fields of data-dependent research will be facilitated by the availability of larger memory systems at near commodity prices," says , PSC Scientific Director, Michael J. Levine, . "Having large amounts of data in directly-addressable memory avoids very time-consuming disk input/output and allows a much more productive programming paradigm."

Researchers' calculations are limited not only by processor speed but also by access to and efficient use of vast amounts of data. And they prefer the familiar programming and runtime environment they are used to with workstations, which Numascale enables. Application areas that require very large memories include natural language processing, multi-organism genomics and quantum chemistry.

"We see the collaboration with Pittsburgh Supercomputing Center as an important milestone for utilizing NumaConnectTM for a number of applications that have previously been limited by inferior memory capacity in standard servers," says Einar Rustad, CTO and co-founder of Numascale. "PSC's unique expertise will strengthen our focus on applications that are key to advances in major scientific fields and help us to widen the market for Numascale."

The collaboration between PSC and Numascale seeks to leverage PSC's unique and extensive experience with very large-memory computing systems and Numascale's NumaConnect memory technology to produce systems capable of handling such large data volumes without memory-retrieval lags. NumaConnect uses commodity servers as building blocks to provide memory capacities and retrieval speeds currently only available through high-end and enterprise-class systems. PSC's application specialists will work with Numascale engineers and application programmers to find ways the two organizations' experience and expertise can be combined synergistically.

BLACKLIGHT RESEARCH SPURS CHANGE IN STOCK EXCHANGE RULES

Findings on the effects of "odd lot" trades on the financial markets, using computations on PSC's Blacklight, have led the New York Stock Exchange, the NASDAQ Stock Market and the Financial Industry Regulatory Authority Inc. to redefine how the industry tracks small stock trades.

Previously, odd lots—trades of 100 or fewer shares—did not have to be reported to regulators. The rationale was that these trades involved small investors who were unlikely to affect the larger market significantly. But recent volatility in the markets, driven by automated small trades that occur far faster than any human can think, called that assumption into question.

In an upcoming paper in *The Journal of Finance*, Mao Ye and Chen Yao of University of Illinois, Urbana-Champaign, and Maureen O'Hara, Cornell University, report that odd lots are playing an increasingly important role in the wider behavior of the markets. The researchers used Blacklight and the San Diego Supercomputer Center's Gordon to analyze market data for the effects of odd-lot trading.

"For every 100 trades of Google, 52 to 53 of them" are in the form of odd lots, Ye observes. "There are more missing trades than trades you can see. In terms of volume, more than 20 percent of the trading volume [among all stocks] is missing" in the official count.

The widely held suspicion is that the largest and most sophisticated traders are using automated trading in odd lots to hide their activities from other traders. In any case, the researchers showed that including the odd lots significantly alters our understanding of the markets. Partly in response to this research, in June 2013 the market authorities agreed to a plan to require all trades, of as few as one share, to be reported starting in October.

"In the U.S., they care a lot about the transparency of the market," Ye explains. The new rule change will remove "a kind of darkness we cannot see and that we never realized was there."

PSC PATENTS SOFTWARE FOR PROTECTING SUPERCOMPUTING RESULTS AGAINST SYSTEM FAILURES

PSC scientists have patented ZEST, a piece of software that takes a rapid "snapshot" of a supercomputer's calculations as it works. ZEST greatly speeds the ability to store complex calculations as a hedge against a system failure, saving precious supercomputing time and slowing calculations down far less than current methods. PSC co-inventors of ZEST included Paul Nowoczynski, Jason Sommerfield, Nathan Stone, and Jared Yanovich.

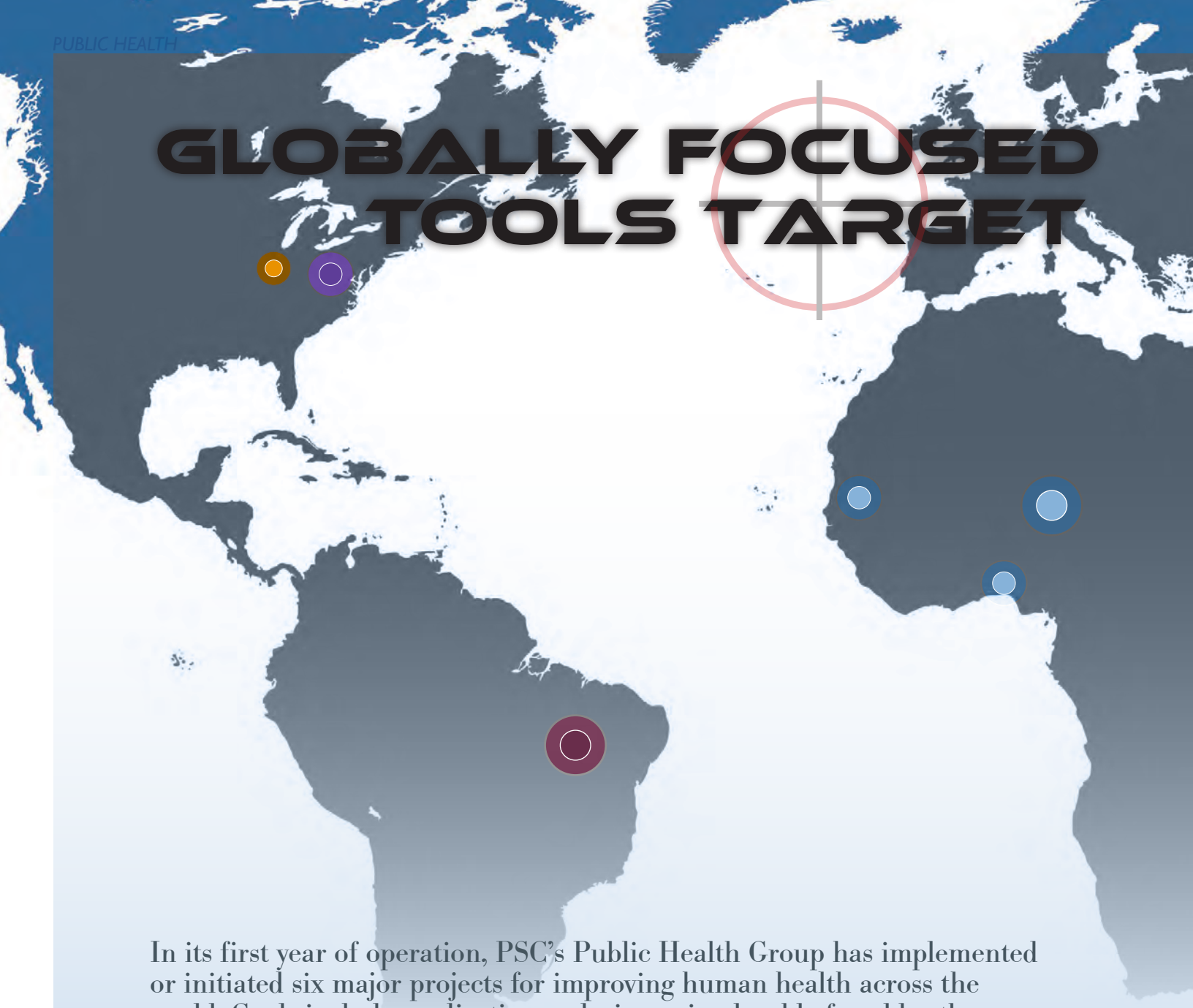
Just as we all hit "save" periodically to avoid losing our work in case of a crash, scientists carrying out vast computations such as those required for detailed weather predictions or earthquake science need to periodically store—"checkpoint"—the machine's computational state.

The problem, according to J. Ray Scott, director of systems and operations at PSC, is that retrieving and storing these data takes time away from calculation, which is carefully rationed to researchers using highly in-demand supercomputers. In fact, he adds, over the last seven years the memory available in the largest machines has increased about 25-fold, while the capacity for retrieving that memory has increased only about four-fold.

"If you have a large job, checkpointing the run often means writing out tens of terabytes of data"—enough to fill about a thousand new iPads, Scott says. "This takes a nontrivial amount of time. The whole time you're doing the checkpoint, you're not using the computer."

The ZEST software works by tightly managing the supercomputer's disk drives, continuously routing checkpoint storage to disks that aren't being used for computation. ZEST has demonstrated 90 percent of the theoretical maximum efficiency of writing data to drives; currently available commercial systems have efficiencies of 25 percent or less.

GLOBALLY FOCUSED TOOLS TARGET



In its first year of operation, PSC's Public Health Group has implemented or initiated six major projects for improving human health across the world. Goals include eradicating malaria, an implacable foe older than the human race; creating computer models of individual actors that completely test every detail of disease transmission and joining software and stakeholders to allow experts in public health to leverage each other's expertise.

- *Location: U. S.*—Two grants to PSC public health researchers and colleagues elsewhere support the MIDAS project's mission to advise U. S. decision makers on public health policy. The *FRED* agent-based modeling software includes independently acting electronic avatars for 300 million individuals to simulate human-to-human disease transmission. *GAIA* affords geospatial imaging to help link geographic obstacles and opportunities to health decision making.

- *Location: Brazil, Thailand, Australia*—The *CLARA* agent-based model simulates the interactions between humans and mosquitos, to test the effects of new treatments, preventive measures and mutations on the spread of dengue fever.

HEALTH PROBLEMS WORLDWIDE

Location: Benin, India, Niger, Senegal, Thailand, Vietnam—The *HERMES* Logistics Modeling Team at the University of Pittsburgh and PSC has dynamically modeled the vaccine delivery system in the West African nation of Niger. In the process, they discovered that transportation of vaccines is an under-appreciated bottleneck deserving more attention and support. In addition to Niger, *HERMES* team members have begun working with the governments of Benin, India, Senegal, Thailand and Vietnam to produce individual analyses of those countries' vaccine supply chains.

Location: Solomon Islands, Kenya—In collaboration with the University of Notre Dame, PSC is creating the electronic infrastructure supporting *VECNet*, a Web-based clearinghouse of ideas and methods that will enable researchers, clinicians, aid agency personnel and government decision makers to share data and test tools for fighting malaria.

Location: Pittsburgh—When deployed, *Apollo* will create a Web-based electronic bridge that allows otherwise incompatible public health models to communicate with each other. One early goal will be to allow *FRED*'s agent-based model of human disease transmission in the U. S. to integrate the geospatial data available through *GAIA*.

Niger's vaccine supply chain superimposed on a map of the country.

Concentration

PSC Blacklight, Data Supercell Enable Non-Human Primate Reference Transcriptome Resource to Support Study of Genes in Our Closest Relatives

In the card game "Concentration," you place the 52 cards in a deck face down on a table. You turn one over; then you turn over another, with the intent of matching the numbers. If they don't match, you turn them face down again. Then you repeat, trying to find the matches.

Memory is paramount. When you see a "6" card, you have to remember where you last saw that number.

Now imagine a game of Concentration in which the task not only involves 3 billion cards, but also the multiple ways they can be strung together to make winning poker hands. And imagine what kind of memory you'd need to find the matches.

Thanks in part to PSC Blacklight's best-in-world shared-memory and the Data Supercell's ability to store and move huge amounts of data in a fluid and accessible way (see Technical Note, p. 21), the laboratory of Christopher Mason, assistant professor in the departments of Physiology and Biophysics and the Institute for Computational Biomedicine, Weil Cornell Medical College, and colleagues have spearheaded the first repository of the active genes in 13 nonhuman primates. The effort has been led by Lenore Pipes, an NSF graduate research fellow and student of Mason and Adam Siepel, associate professor of computational biology, Cornell University.

Blacklight is a SGI UV shared-memory system; the Data Supercell, PSC's archival system, is a low-cost, high-bandwidth, high-capacity and highly-reliable data management system.

Reported in a January 2013 *Nucleic Acids Research* paper, the Non-Human Primate Reference Transcriptome Resource (NHPRTR) provides an electronic infrastructure to support researchers who are sequencing, comparing, and trying to understand genes in mankind's closest relatives.

GETTING THE GEOGRAPHY DOWN

Understanding the genomes of the nonhuman primates—great apes such as chimpanzees, old world monkeys, new world monkeys and more primitive prosimians, such as lemurs—is important for understanding ourselves in health and disease, Mason explains.

"Nonhuman primates are widely used models in pharmaceutical research," he says. "Also, understanding their genomes allows us to answer evolutionary questions about how the human genome came to be."

But the state of the field in primate genome research is uneven, depending on which species you look at.

"Not all primates' genomes have been sequenced," Mason says. Chimps and rhesus monkeys have been sequenced—but not enough times to ensure good proofreading. "Even for those that have been sequenced, there is an incomplete or poor annotation of what genes are present in these species."

Annotation is critical. If the genome were a map, for example, the DNA sequence would represent lines

for the roads and circles for the cities. Annotation is the process of putting labels on that map.

"If you took a normal map but removed all the names of the cities, you wouldn't know where Philadelphia is, where Pittsburgh is," Mason says. "You would have no sense of geography. Annotation lets you navigate the human and primate genomes in the same way that an annotated map lets you navigate the U.S."

UNCOVERING THE CARDS

The initial work for the NHPRTR consisted of obtaining the sequences for the active genes in 13 primate genomes.

In the cell, DNA is the master copy of the genetic material, encoding the blueprints for making the cell's components in a series of bases: A, T, G and C. In order to express a gene, the cell copies its DNA sequence to RNA, a molecule closely related to DNA. The cell translates some of these RNA copies into proteins, the main actors in the cell. Other RNAs carry out specific functions on their own. The process of copying a gene's DNA code into RNA is called transcription. The transcriptome is the collection of RNAs that are being expressed in an organism's living tissues.

The process of reading the RNA code requires fairly short segments—about 100 bases, for optimum efficiency. Because the genome is so much larger than that, researchers must first cut it into small, overlapping fragments. Once they have the sequences of these fragments—about 600 million of them for a typical species, though some of the primate group's analyses looked at as many as 3 billion—the researchers can then reconstruct the entire transcriptome sequence by matching where the segments overlap. It's much like playing a game of Concentration that follows all the ways that 3 billion cards could make a winning hand.

But it gets harder. Many of the sequences are nearly, but not completely, identical. In addition, inevitably there are some errors in the sequence that have to be corrected, by covering each bit of the sequence multiple times. It's hard to piece imperfect, redundant and nearly identical bits together in the proper order. In order to get it right, the scientists must construct a chart of possible ways of stringing them together, testing each possibility in turn. These charts are called De Bruijn graphs.

BLACKLIGHT AND DATA SUPERCELL: ALL ABOUT THE MEMORY

Blacklight speeded the calculation behind the matching considerably, Pipes says. A traditional, "distributed memory" supercomputer would have solved the problem by raw speed only, essentially uncovering each "card"—possibility in the De Bruijn graph—one at a time, then turning it face down to check another. Blacklight's massive shared-memory, though, made that unnecessary—the machine was able to keep many possibilities in its memory at once, allowing for far more rapid matching.

PSC's Data Supercell (see Technical Note, p. 21) also allowed the researchers to use their massive amounts of data efficiently, making it available in large chunks with minimal retrieval delay.

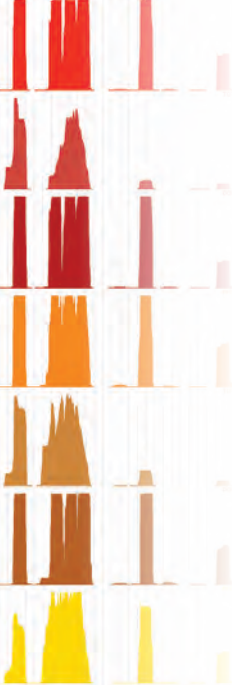
Running the problem on Blacklight required PSC staff to work with the developers of the De Bruijn graph software, Trinity, to optimize its performance on the new machine.

"The primate NHPRTR study was the biggest thing anyone had tried to do with Trinity," says Philip Blood, PSC senior scientific specialist, who helped the researchers get Trinity running on Blacklight. Getting the program to work on the new platform required painstaking trial and error.

ANNOTATION: TABLE OF CONTENTS FOR EXPRESSED GENES

Some primate genomes have been sequenced already. More are likely to come as the technology becomes quicker, easier, and cheaper.

"Sequencing DNA is really only the first step in understanding a genome," Mason says. "You need to know what's expressed, what's active" from a DNA sequence, in the RNA, to make sense of how an ani-



mal's genome works, he adds. "Our transcriptome maps create the first catalog of functional, active elements in a given genome. That is the first essential step of delineating the molecular recipe that defines the synthesis of the entire organism, from one cell to the trillions of cells in an adult."

The transcriptome contains the sequences of all the active genes in a given cell or tissue. By comparing the sequences in the transcriptome to those in the genome, researchers can tell which genes are active. It's a particularly important question in understanding why humans and chimpanzees are different. Both species' DNA is very similar—96 percent of the sequences are identical. Because of that many researchers suspect that they may express different genes at different times in specific tissues, painting a very different picture out of an almost identical palette of colors. This is the sort of question that the NHPRTR will help primate researchers to tackle, Pipes says.

At the moment, the NHPRTR represents 13 key primate species identified by researchers in the field. In the next stage of the research, they will focus on deep sequencing of matching tissues from various individuals, including the molecular characterization of the various brain regions between these animals. They'll also look at the evolution of gene structure, alternative splicing and the catalog of species-specific genes. Taken together, these data inform not only the functional map of genes for these primates, but ultimately help answer the age-old question at a genetic level, "What makes us human?"

Technical Note

PSC and Galaxy: Enabling Big Data to Scale Even Bigger

One of the strengths of PSC's Data Supercell (DSC) is that, by making Terabyte-scale data rapidly accessible, it enables projects such as the Non-Human Primate Reference Transcriptome Resource to store large amounts of data in a rapidly accessible way (see p. 18). In another effort, with the Galaxy Project, the system's speed and volume enables an already large-scale research venture to grow to meet its users' needs.

Galaxy was conceived as a way for biologists with limited programming and information technology background to easily make use of tools for sequencing and analyzing DNA and RNA. A team led by Anton Nekrutenko, associate professor of biochemistry and molecular biology at Penn State University and James Taylor, associate professor, departments of Biology and Math & Computer Science at Emory University, had written the freeware Galaxy program as an easily learned interface that helped these researchers organize their workflow to make optimum use of high-performance computing tools. In addition, Galaxy is "domain agnostic"—it can be adapted to just about any scientific field requiring high performance computing and complex analysis of large amounts of data.

Another virtue of Galaxy is that it transparently saves all the metadata behind an analysis, says Nekrutenko. This allows scientific reproduction of results by other researchers.

Galaxy, however, was becoming a victim of its own success. It proved so popular that within just a few years its user community had grown to about 33,000, not including public and private instances users had installed in their own systems. With that volume of users, inevitably the system's administrators had to impose quotas and couldn't allow some resource-intensive potential applications.

In addition, some increasingly important types of analyses simply weren't possible with the computing resources available to Galaxy. De novo genome or transcriptome assembly, in which researchers piece together small bits of DNA or RNA, respectively, to decode an organism's full genome or its complement of active genes, require large amounts of computer memory that outstripped what local machines could do.

The ability of Galaxy to keep pace with is users, let alone to grow further, was in question.

PSC offered a solution that combined the center's hardware and software resources to overcome these limitations. The resulting combination of the DSC's storage capacity, PSC Blacklight's best-in-world shared-memory, a direct 10-gigabyte/sec Ethernet connection between PSC and Penn State and PSC's SLASH2 software, which allows them to communicate rapidly with computers at Penn State, have allowed large-memory users to submit their jobs to Galaxy and have them run transparently on Blacklight.

In addition to allowing more memory-intensive jobs easy access to Blacklight, the integration of Galaxy with PSC will enable users to employ memory-intensive sequencing tools, says Phil Blood, PSC senior scientific specialist and liaison to Galaxy. Programs such as Velvet and Trinity—for brute-force sequencing of DNA and RNA, respectively—are now available to users.

Mind the Gap

In 1996, Ricardo González Mendez (not pictured) decided to revamp his skills and expertise in bioinformatics—using advanced computing techniques on biological problems. In the process, he learned something alarming: American biology education was in danger of becoming a two-class system.

A gap is developing, the University of Puerto Rico (UPR) School of Medicine professor realized. The top-tier institutions understand that bioinformatics will soon be a job requirement in much of biology. And they have expended their considerable resources to create bioinformatics classes, degree programs and research centers. Among the other institutions, many realized as well that a bioinformatics crisis was coming. Their students were in danger of being left behind, whether they wanted to pursue academic research, industrial positions or even teach. But the schools had neither the expertise nor the resources to respond.

MARC Program Helps Minority-Serving Institutions Prepare Students for 21st Century Biology Careers

"There is a total disconnect between the top-tier research schools and the minority or not-so-rich schools," Gonzalez warns.

Today's biologists are generating incredible amounts of data. The European Bioinformatics Institute alone, for example, now stores 20 petabytes of life sciences data—enough to fill nearly a quarter of a million top-line iPads. Understanding that much data can only be accomplished with computers.

"People are starting to come to the realization that either they modernize their skills, or they're not going to get more funding," Gonzalez says.

Gonzalez enlisted the then-director of PSC's biomedical initiative, David Deerfield, who with PSC's Hugh Nicholas and Alex Ropelewski wrote a proposal for a 2001 National Institutes of Health Minority Access to Research Careers (MARC) grant to help students at minority-serving institutions study bioinformatics.

Gonzalez was the PSC MARC program's first faculty liaison. After Deerfield's tragic death in 2006, he took on an expanded role as co-principal investigator with Nicholas.

Initially, the program focused on helping institutions establish a single bioinformatics course on campus. "The focus of the current grant is on working with five minority-serving [partner] institutions to get concentrations in bioinformatics on their campuses," says Ropelewski. "It's basically a multi-course work series that's established officially, so that someone can minor in bioinformatics, for example." The program fills gaps in students' knowledge base—teaching biologists about computers and computer scientists about biology.

PSC's MARC program has enjoyed a number of successes. The students are "publishing in really good journals, getting funded from various agencies, and ... going on to graduate programs, post-doctoral fellowships and positions in industry and government that are very good," Gonzalez says. "We are not as big as some [of the top] programs, but we produce the same kind of quality."

One of PSC MARC's most exciting facets is a 10-week summer program, in which students from participating institutions come to PSC to carry out bioinformatics research projects. Here we present a sampling of this year's students and their efforts.

PUTTING THE "TECH" IN BIOTECH

Michael Thompson loves the gadgets.

"To be honest, I really love technology," says the incoming Jackson State University senior. "I'm just fascinated by it."

As a freshman Thompson wasn't sure what he wanted to study. But he'd scored high in biology in the Mississippi state high school tests, and when Raphael Isokpehi, director of Jackson State's Center for Bioinformatics and Computational Biology, invited him to visit Isokpehi's lab he decided to check it out.

"There were a lot of computers but I didn't notice any microscopes or anything like that," Thompson says. "I thought, 'That's weird.'" That's where, for the first time, Thompson found out that he could do biology *and* engage in his love of technology.

At the MARC summer workshop, Thompson continued his project from Isokpehi's lab, studying the universal stress proteins (USPs) in the *Clostridia* bacteria. USPs are an important part of an organism's defense against stresses such as antibiotics. Because of that, they're an important target for treating food poisoning, colitis, tetanus and other diseases caused by *Clostridia*.

"Being able to have Alex, Hugh and Pallavi [Ishwad, PSC's Education Program director] guide me in the right direction has been amazing," Thompson says. "I feel like I can take a lot back home and teach others."

GRABBING THE BRASS RING

People had plenty of advice for Tevin Reed, an incoming senior at North Carolina Agricultural and Technical State University, about what *not* to major in. His band director warned about the grim employment prospects for a brass musician. Reed's sister, an information technology major, always seemed to need another expensive tool for her projects. But in computer science, she told him, "As long as you have your laptop you can always do your homework."

Reed found he loved computer science. "I'm not going to say I was the best at it, but I could actually understand what the teachers were saying on the first day," he says.

Reed's MARC project was to use the wxWidgets library to make the open source GeneDoc Windows program work on any operating system. Hugh Nicholas, the late David Deerfield and Alex Ropelewski developed program specifications and Nicholas' son Karl coded GeneDoc, which visualizes, highlights and rigorously compares DNA, RNA and protein sequence alignments. Reed hopes to help biologists get more consistent computational results no matter what kind of computer they use.

"It was very helpful" having the program's authors available in working out the inevitable glitches, Reed says. "Once I finish it, I'll talk with Alex and Hugh and see how they want to distribute it."

GETTING A BETTER VANTAGE POINT

From atop Ingrid Montes-Rodriguez's grandparents' house in Ciales, in the mountains of Puerto Rico, you can see both the Atlantic Ocean to the north and the Caribbean Sea to the south.

"It's really beautiful," says Montes-Rodriguez, a PhD student in Juan López Garriga's chemistry lab at the University of Puerto Rico, Mayagüez, who is currently doing her research work at the UPR Medical Sciences Campus. When the pressure of grad school gets to her, she heads for Ciales and her family.

Her dad is a master of "tough love" advice: "He just says, 'Well, you have to do it! What are you going to do, cry?'"

Thanks in part to that advice—and a PSC MARC summer project with Graham Hatfull's lab at the University of Pittsburgh—Montes-Rodriguez has begun decoding the genetic material of the clam *L. pectinata*, a major focus of López Garriga's team.

L. pectinata, which grows in mudflats throughout the Caribbean, survives levels of hydrogen sulfide that would kill most animals. It does this in part by producing a unique kind of hemoglobin, which attaches to that "rotten egg" chemical instead of oxygen.

Montes-Rodriguez hopes that comparing the genome of the clam with the DNA of other species will help explain how the unique hemoglobin evolved, and what other protective mechanisms the species has developed.

PSC MARC Program Participating Institutions

developing bioinformatics concentration program	offering bioinformatics course
Jackson State University	Howard University
Johnson C. Smith University	Morgan State University
North Carolina A&T University	North Carolina Central University
University of Puerto Rico, Mayagüez	Universidad Metropolitana, Puerto Rico
Tennessee State University	University of Puerto Rico, Medical Science Campus
	University of Texas, El Paso
	University of Texas, San Antonio

Summer Job

Not all the success stories in this summer's MARC program involve MARC students.

Jonathan Strickland, a recent high school graduate of the Pittsburgh Science and Technology Academy, knew that he wanted to work with computers. It's probably fair to say, though, that after a senior project at PSC he's aiming considerably higher in the industry—and it started with this year's summer employment.

"I kind of pictured having a summer job at Arby's," he says. But the now-University of Arizona Honors College freshman and full scholarship recipient found himself on a different track. Working with PSC's Alex Ropelewski, he did a project analyzing what factors affect the RNA sequence assembly program Trinity's performance on supercomputers (see p. 18).

"I guess I did a good job, because Mr. Ropelewski asked me if I wanted to work at PSC this summer" helping with the MARC program.

So Strickland helped keep the MARC workshop running smoothly. He did some purely gopher tasks. But he also set up user accounts for the MARC students. He even helped out with the Python programming language class, which he himself learned during his senior project experience.

"Working at PSC is great," he says. "I'm actually doing stuff that's meaningful."

Needle in a Needlestack

PSC's Sherlock Supercharges Next-Generation Search Tool

We've all had the experience. Enter what seems to be a straightforward search term, hit the "return" key and Katie-bar-the-door.

Off-target links. Duplicative links. Unmentionable links. Often, too many links to read.

There's no doubt that Internet searches have revolutionized our ability to locate information. They've essentially solved the "needle in a haystack" problem. But they've created another problem. Search results brim with almost-appropriate documents that block our view of what we need. What we're left with, essentially, is the problem of finding a particular needle in a *needlestack*.

A new collaboration between PSC and CFL Software promises to turn that problem around, returning less—but far more relevant—information. CFL, which specializes in linguistic document forensics, is pairing PSC's Sherlock with their program CFL Discover. The intent is to improve our ability to comb through larger, more complex and chaotic data sources to provide a workably small number of findings that connect to a search term in productive—and in many cases, unexpectedly productive—ways.

The combination of Sherlock and CFL Discover promises far more targeted searches of unstructured datasets. Shown left is a "spider graph" of an arbitrary Twitter account, a representation typical in such graph analytics.

TRACKING IT DOWN WITH SHERLOCK

"The fundamental challenge with searching large-scale, complex knowledge bases such as unstructured text documents is that investigating them in small parallel pieces, the way a traditional supercomputer would, is very difficult," says Nick Nystrom, director of strategic applications at PSC. "The dense connections between the information in the documents make it impossible to partition the data effectively between processors on traditional supercomputers or clusters."

This leads to an extreme case of what is called the "memory wall," in which the computer spends most of its time waiting for data to be retrieved from memory rather than calculating results. Sherlock—a modified YarcData Urika™ graph analytics appliance launched in February 2013—has particular strengths for such searches that stem from its architecture. YarcData is a subsidiary of Cray, Inc.

"Sherlock combines purpose-built processors that execute 128 threads, each with a custom network to support graph analytics," Nystrom says. This enables the device to carry out multiple overlapping calculations at once without hitting the memory wall.

Sherlock is supported by a $1.2-million grant from the Strategic Technologies for Cyberinfrastructure program of the National Science Foundation.

NEW SOFTWARE

By making the CFL Discover software available to researchers on Sherlock, the strategic partnership intends to redefine what we expect from search technology.

"This is a new venture both in terms of scale and speed in searching for information," says David Woolls, CEO of CFL Software, which specializes in linguistic document forensics. While many users may not be aware of it, search engines don't completely search all the text in the entire Web—that would take far too long. Instead, they search "metadata" that have been added to those documents. Some metadata must be added by humans, a process that's time-intensive and incomplete. The result is a search process that's inexact.

"Search engines start with a few words and return a list of documents which contain them," Woolls adds. "CFL Discover starts with one or more of those documents and reads them for you, shows you the terminology that is shared and gives immediate access to the passages of particular interest to you."

LEVERAGING CFL DISCOVER WITH PSC RESOURCES

CFL's new work with Sherlock will explore a substantial portion of the U.S. Patent database, in addition to the full data of Wikipedia.

"PSC's role in the partnership is to couple the unique analytic capability of Sherlock running CFL Discover with hosting massive datasets on PSC's Data Supercell," Nystrom says. "This will help expand text analytics to unprecedented, interdisciplinary use cases."

Potential examples include analyzing recent events from news and social media sources, extracting deeper insights from sets of publications, and supporting computational history and culturomics—the quantitative study of cultural phenomena by analyzing large volumes of written records.

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

PSC gratefully acknowledges significant support from the following:

The Commonwealth of Pennsylvania
The National Science Foundation
The National Institutes of Health
The National Energy Technology Laboratory
The National Oceanographic and Atmospheric Administration
The National Archives and Records Administration
The U. S. Department of Defense
The U. S. Department of Energy
D. E. Shaw Research
Cisco Systems, Inc.
Cray Inc.
The Grable Foundation
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

PHOTOGRAPHY: Tim Kaulen, Photography & Graphic Services at Carnegie Mellon University.
GRAPHICS/VISUALIZATIONS: Thanks to the researchers, PSC scientists and Shandra Williams.
COVER GRAPHIC: The combination of Sherlock and CFL Discover promises far more targeted searches of unstructured datasets. One tool Sherlock generates for investigating relationships is the spider graph. Shown here is a representation of the connections in an arbitrary Twitter account.
PRINTING: Hoechstetter Printing

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

What can you do with a high performance graph analytics appliance like Urika™?

Our customers are using it to:

- Accelerate drug discovery
- Investigate fraud
- Enhance risk and compliance efforts
- Identify new customer insights
- Improve network analysis

We help you analyze graphs at scale so you can validate new hypotheses faster — to get to valuable insights.

YarcData™
Getting to **Eureka!** faster™

www.YarcData.com

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