SAN DIEGO SUPERCOMPUTER CENTER at UC SAN DIEGO

DATA to DISCOVERY

Pioneering a New Era of Scientific Investigation

SDSC

ANNUAL REPORT 2013

SAN DIEGO SUPERCOMPUTER CENTER

As an Organized Research Unit of UC San Diego, SDSC is considered a leader in data-intensive computing and all aspects of "Big Data," which includes data integration, performance modeling, data mining, software development, workflow automation, and more. SDSC supports hundreds of multidisciplinary programs spanning a wide variety of domains, from earth sciences and biology to astrophysics, bioinformatics, and health IT. With its two newest supercomputer systems, *Trestles* and *Gordon*, SDSC is a partner in XSEDE (Extreme Science and Engineering Discovery Environment), the most advanced collection of integrated digital resources and services in the world.

SDSC INFORMATION
Michael L. Norman, Director

San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive MC 0505
La Jolla, CA 92093-0505
Phone: 858-534-5000
Fax: 858-534-5152

info@sdsc.edu
www.sdsc.edu

Warren Froelich
Director, External Relations wfroelich@sdsc.edu
858-822-3622

DATA to DISCOVERY
SDSC Annual Report 2013
(PDF version available online at the SDSC website)

EDITOR:
Warren Froelich

CO-EDITOR:
Jan Zverina
Director, Communications and Media Relations

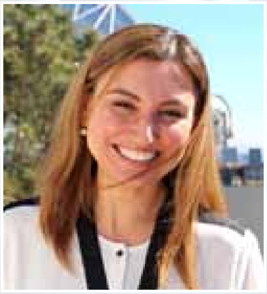
CONTRIBUTORS:
Diane Baxter, Cassie Ferguson, Warren Froelich,
Ron Hawkins, Alma Palazzolo, Susan Rathbun,
Mahidhar Tatineni, Jan Zverina

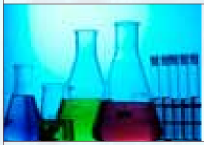
CREATIVE DIRECTOR:
Ben Tolo

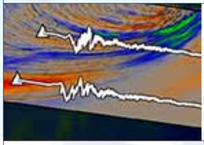
DESIGN:
Ben Tolo, Monika Braun

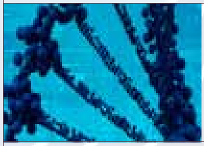
Any opinion, conclusions, or recommendations in this publication are those of the author(s) and do not necessarily reflect the views of NSF, other funding organizations, SDSC, or UC San Diego. All brand names and product names are trademarks or registered trademarks of their respective holders.

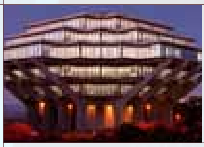
© 2014 The Regents of the University of California











DIRECTOR’S LETTER

SDSC’s Response to the “BIG DATA” CHALLENGE

During the past couple of years, our planet began confronting what many view as a tidal wave of information stemming from our academic centers, commercial laboratories, government scientists, and observational tools such as satellites, oceanographic sensors, astronomical observatories, and personal websites. The term “Big Data”, used widely to describe this phenomenon, suddenly became a major topic across a wide spectrum of interests and disciplines which began seeking ways to tame, harness, and otherwise glean knowledge from this ever-widening deluge of data.

Thought leaders from the highest levels of the federal government to board chairmen at the world’s leading private and public enterprises have met in workshops, seminars, and other local and national gatherings to better understand the concept of Big Data, initially, at least, to separate hype from reality. These leaders hope the ongoing Big Data conversation yields better narratives and/or visualizations needed to help explain events of importance to science and society, from the seeming fickleness of the marketplace to how virulent diseases spread from neighborhood to neighborhood, city to city, and beyond.

Our nation’s research universities, a traditional source of innovation and discovery, also have been brought into the discussion by the business community and government leaders. Here at UC San Diego we’re being challenged to develop new tools to better mine, model, manage, and otherwise analyze all this data; to build effective methods to store and curate huge volumes of information over long periods of time; and to develop curricula to help educate and train experts to fill a rapidly growing need for experts in all things relevant to Big Data. The term “data science and engineering” has emerged as academia’s answer to the Big Data need, much the way “computational science and engineering” programs sprung up a couple of decades ago in response to an advancement known as supercomputing.

In this year’s annual report, we describe an innovative program at SDSC called the Institute for Data Science and Engineering (IDSE), which provides a roadmap for UC San Diego’s response to the Big Data challenge. As outlined, IDSE will host and coordinate undergraduate and graduate education and training of Big Data experts at UC San Diego, while also providing a focal point for research collaborations across campus, particularly for its recently unveiled strategic themes that include: understanding and protecting the planet, enriching human life and society, exploring the basis for human knowledge and creativity, and understanding cultures and improving societies.

IDSE also will offer an infrastructure to support Big Data such as: scientific simulation and visualization, data modeling and integration, database design and implementation, scientific workflow automation, data mining and predictive analytics, and management of protected data. Over the course of its nearly three-decade history, SDSC has fostered research collaborations and partnerships across a variety of disciplines and departments at UC San Diego and beyond, and we envision IDSE as a lightning rod for new efforts to advance this campus' national influence in data science and engineering.

The evolution of data science and engineering complements SDSC's more traditional focus on advancing computational science and engineering. Today, both scientific methodologies are fundamental components of the same tool kit that researchers need to discover new concepts and advance innovative technologies. For example, next-generation sequencing of DNA and RNA—the foundation for the advancing era of "personalized medicine"—is creating a deluge of data that requires increasingly powerful supercomputers to process, and the need to develop complex software to more rapidly and efficiently analyze. SDSC is leveraging its Big Data resources including the *Gordon* data-intensive supercomputer to support efforts in genomics and bioinformatics, working with researchers at UC San Diego and UC, neighboring non-profit research centers on the Torrey Pines mesa and beyond, and California biotech companies.

This year's annual report shines a spotlight on SDSC's vast array of Big Data and computational science capabilities that includes its supercomputing resources, advanced networking and data storage infrastructure, along with brief descriptions of this Center's local impact, statewide influence, and national reach in education, training, and research. The report also excerpts significant research advances made by, or with the help of, SDSC staff and resources during the past year that include new academic collaborations and industry partnerships.

Over its nearly three-decade history, SDSC has garnered a national reputation for excellence in data and supercomputing and, partly as testimonial to this reputation, was awarded a $12 million grant during 2013 from the National Science Foundation (NSF) to build a next-generation supercomputer capable of one quadrillion arithmetic calculations per second. This petascale data-intensive supercomputer is fittingly called *Comet*, connoting its mission of catering to what's being called the "long tail of science," whose goal is to address the needs of a large number of small- to medium-sized computationally based research projects nationwide. In essence, *Comet* is all about high-performance computing for the 99 percent, and is designed to deliver a significantly increased level of computing capacity and customizability to support data-enabled science and engineering for the nation. *Comet* will have all the features that made SDSC's *Trestles* computer cluster so popular with users, and will be particularly well-suited to science gateways that serve large communities of users, such as those new to XSEDE (NSF's eXtreme Science and Engineering Discovery Environment) which comprises the most advanced collection of integrated digital resources and services in the world.

We realize that the computational and data science resources at SDSC are only as good as the people who use them. A critical mass of expertise across numerous specialties support and integrate these various components into a holistic working environment. These individuals include scientists with Ph.D degrees in computer or domain sciences who, through their collaborations, have become cross-trained in a range of computational/domain sciences to help solve problems in the latter. An emerging description of such people is "pi-shaped," with one leg in information technology, one leg in a scientific domain, and a bar across the top indicating the know-how to bridge the two culturally distinct communities.

SDSC is proud of its cadre of "pi-shaped" researchers and, for the first time, we have chosen SDSC's Ilkay Altintas as SDSC's "Pi Person of the Year." Please read about Dr. Altintas' accomplishments later in this annual report, and learn more about how SDSC is shaping its future, and those of its partners and collaborators, in this exciting new era in science and technology.

Michael L. Norman
SDSC Director

Ilkay Altintas

SDSC π Person of the Year

Keeping *the* Flow *in* Scientific Computing

Ilkay Altintas, deputy coordinator for research and director of SDSC's Scientific Workflow Automation Technologies Laboratory, was recently named SDSC's first "Pi Person of the Year." Named after the π symbol, the award recognizes researchers who, as collaborators and innovators in applied research and development, have one 'leg' in one or more science domains, and the other in cyberinfrastructure technologies. Altintas makes sure those researchers can focus their resources on their science rather than solving workflow issues or other computational problems. Altintas has been exploring scientific workflows for over a decade, and is one of the founders of the Kepler collaboration, a project that provides researchers the means to access, arrange, and share data and workflows via a common interface.

What is a scientific workflow and Kepler in particular?

Altintas: A scientific workflow is a series of computational steps that scientists use to generate results. That may involve accessing multiple applications and databases, and processing the data using computationally intensive jobs on high-performance clusters. Kepler is an open source, community-based scientific workflow application that helps users share and reuse data, workflows, and components developed by the scientific community to address common needs. It really started as a grassroots effort just as the compute grid started taking off, and over the past decade has evolved as integrating software has become increasingly difficult. It is named Kepler, after the Ptolemy software on which it is built.

How do you make the scientific community aware of Kepler?

Altintas: In the early days, we had to pitch the tools to communities, go to the researchers, and explain how it would fit in with their work. These days we still recruit, but there is a new generation of scientists that find workflows critical to the success of their work even though they might not be programmers or have the money to hire one. When they become interested in using workflows they find us.

What science domains use Kepler?

Altintas: Kepler has been adopted by researchers in all scientific fields—geophysics, astrophysics, computational biology, ecology, and more. As the need for workflows has increased, I've been involved in projects ranging from bioinformatics and microbial ecology, to ocean monitoring and wildfire management via sensor networks. They all face similar challenges in software and data integration. UC San Diego has a number of biomedical and cancer projects such as the National Biomedical Computation Resource that use Kepler. A new project, called WIFIRE, will apply workflows to wildfire data, which serves not just scientists, but people in the San Diego community who might be affected by such fires (see page 38).

What do scientific workflows have to do with wildfires?

Altintas: San Diego county has one of the most wired environments in the world with the High Performance Wireless Research and Education Network (HPWREN) remote sensing network, and we have frequent wildfires. The WIFIRE project combines the data from those sensors along with other information such as satellite and weather data, and then applies large-scale computing to signal processing, visualization, modeling, and data assimilation to help monitor and predict wildfire behavior. Scientific workflows help simplify those steps so that we can quickly generate the different kinds of models and image data needed by scientists, policy managers, and firefighters, among others.

What projects would you take on or hardware would you buy if you had unlimited funding?

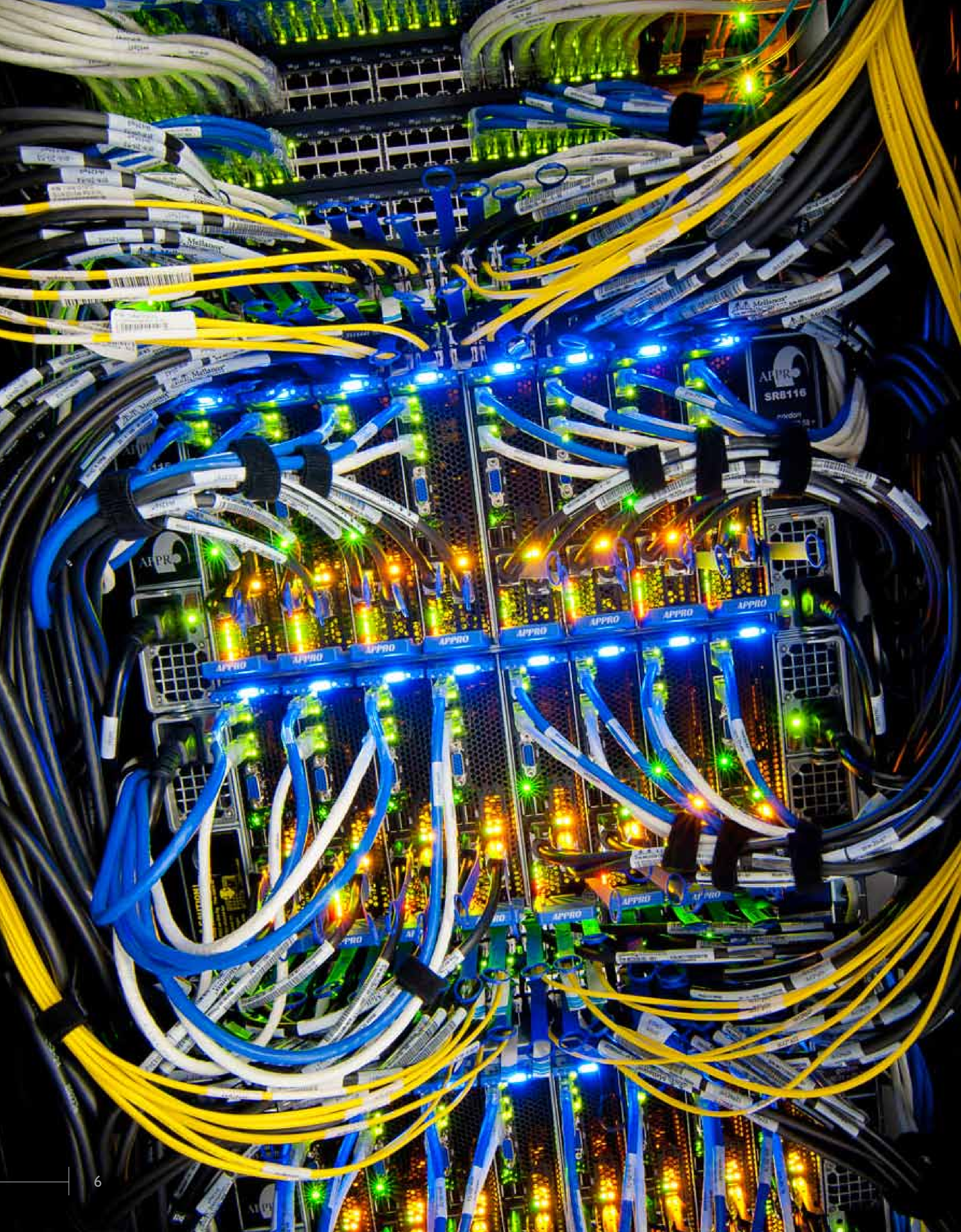
Altintas: It sounds funny to say, but I wouldn't change a thing. I'd try to do what we already do better, more efficiently. Lately I've been interested in "reproducible science." Scientific collaboration just goes forward faster and it is important to capture scientific efforts in a more open way so that it can be reproduced in the future.

You serve on the editorial board of the *Future Generation Computer Systems* journal. What are popular topics?

Altintas: The journal is focused on advanced cyberinfrastructure for large-scale distributed systems, so some of the more popular papers are about workflows, cloud computing, and solving optimization problems.

How did you wind up at SDSC?

Altintas: I had just finished my master's thesis and was working at the Middle East Technical University in Turkey. I saw an interesting job at SDSC working on scientific data management, applied, and wound up moving to San Diego to work as a programmer in 2001. I've been at SDSC ever since, and a lot of good things happened along the way!


APPRO
SR8116
Mellanox

SDSC's BIG DATA CAPABILITIES

For almost 30 years SDSC has strived to provide a unique blend of advanced computing and expertise that serves as a local, state, and national gateway for collaboration, innovation, and education for an era defined and transformed by its information. Guided by this vision, SDSC not only has provided the resources and expertise needed for current research, the Center has sought to anticipate researchers' future requirements through the development and creative use of new tools to advance discovery.

Today, scientific research has entered what is called the "fourth paradigm": data-enabled investigation. That data comes in many forms and from all over the globe—academic and government institutions, commercial enterprise, even observational tools such as satellites and deep-sea sensors. It is streaming in at ever-increasing rates, creating a formidable challenge for researchers to not only extract meaningful value from this data, but to manage, organize, and store massive amounts of information.

SDSC is meeting this multi-faceted challenge by focusing on three key areas of capability: High-Performance Computing and Resources (for rapid data processing); Comprehensive Data Management (for secure and stable storage); and Networking/Connectivity (for rapid data transmission).

Inside SDSC's *Gordon* supercomputer. Image: Erik Jepsen, UC San Diego (left)

SDSC's first supercomputer, the Cray XMP-48, in 1985 (right)

Trtion Shared Compute Cluster
or TSCC servers

HIGH-PERFORMANCE COMPUTING AND RESOURCES

Since its inception in 1985, SDSC has been at the forefront of high-performance computing, developing innovations to make supercomputers more capable, compact, and energy-efficient by orders of magnitude. In addition to smaller clusters dedicated to specific tasks, SDSC currently has three main HPC systems used by researchers at the local, national, and global levels across numerous domains that include both traditional and non-traditional areas when it comes to using advanced computation.

Trestles

High-Productivity Workhorse

Trestles came online in early 2011 to provide researchers not only significant computing capabilities, but to allow them to be more computationally productive. Known throughout the national HPC community as a high-productivity workhorse, *Trestles* is based on the idea that by tailoring a system for the majority of jobs rather than a handful of researchers who run jobs at thousands of core counts, users would be rewarded with high throughput and increased scientific productivity. Today, *Trestles* is recognized as the leading science gateway platform in the National Science Foundation's eXtreme Science and Engineering Discovery Environment (XSEDE) portfolio, with more than 650 users per month run through the popular CIPRES phylogenetics portal alone. *Trestles* users span a wide range of domains, including astronomy, biophysics, climate sciences, computational chemistry, material sciences, and more.

Trestles will be replaced in 2015 by an all-new HPC system called *Comet*, a petascale supercomputer designed to transform advanced scientific computing by expanding access and capacity among traditional as well as non-traditional research domains.

Gordon

Delivering on Data-intensive Demands

Gordon entered production in early 2012 as one of the 50 fastest supercomputers in the world—and the first one to employ massive amounts of flash-based memory, making it hundreds of times faster than conventional HPC systems while having enough bandwidth to handle extremely large datasets. The result of a five-year, $20 million NSF grant, *Gordon* has 300 trillion bytes of flash memory and 64 I/O nodes, making the system ideal for researchers who need to sift through tremendous amounts of data. In effect, *Gordon* is designed to do for scientific research what Google does for web searches.

By the end of 2013, 799 research projects using *Gordon* were awarded among 479 principal investigators. In early 2013, *Gordon* completed its most data-intensive task so far: rapidly processing raw data from almost one billion particle collisions as part of a project to help define the future research agenda for the Large Hadron Collider (LHC). Under a partnership between a team of UC San Diego physicists and the Open Science Grid, *Gordon* provided auxiliary computing capacity by processing massive data sets generated by one of the LHC's two large general-purpose particle detectors used to find the elusive Higgs particle. The around-the-clock data processing run on *Gordon* was completed in about four weeks' time, making the data available for analysis several months ahead of schedule (see page 20).

Gordon, housed in SDSC's Datacenter, is the first supercomputer to employ vast amounts of flash-based memory. Image: Erik Jepsen, UC San Diego.

Rick Wagner is SDSC's High-Performance Computing Systems Manager, responsible for SDSC's Linux-based computing clusters and related systems that support users across a broad range of scientific disciplines.

In addition to leading-edge genomic research (*Gordon* is capable of storing 100,000 entire human genomes) the system is now being used by Arieah Warshel, who won the 2013 Nobel Prize in chemistry for developing detailed computer simulations of complex chemical processes for tasks such as designing new drugs or solar cells. Warshel is using *Gordon* and other SDSC data storage resources exclusively after being allocated 3 million CPU (core processing unit) hours last year.

Triton Shared Computing Cluster (TSCC)

Affordable Computing for Campus and Corporate Needs

In mid-2013, SDSC deployed a new high-performance research computing system called the *Triton Shared Computing Cluster*, or *TSCC*, to serve researchers at UC San Diego and other UC campuses, as well as external academic, non-profit, and corporate users. *TSCC* is operated by SDSC for UC San Diego's Research Cyberinfrastructure (RCI) program. The cluster features state-of-the-art hardware, and a revamped participation model that provides researchers more options for funding their computing-based needs. Primary benefits to participants include gaining access to a much larger resource than they could afford solely for their labs, and having a system that is professionally maintained by full-time staff at SDSC.

Enabled with a wide range of open source and licensed software for science and engineering, *TSCC* is designed to work with the full complement of capabilities at UC San Diego, including high-performance networking and centralized storage systems. Connectivity is available to high-throughput scientific instruments across campus, such as DNA sequencers and mass spectrometers.

Phil Papadopoulos is SDSC's Chief Technology Officer and chief architect behind SDSC's *Data Oasis* storage system. Papadopoulos also is principal investigator for the Prism@UCSD project to build a campus cyberinfrastructure capable of supporting extreme data-intensive communications.

COMPREHENSIVE DATA MANAGEMENT

Developing and operating fast and robust HPC systems is only part of the equation. The ability to provide a secure and stable environment to store data, both for the short and long term, is essential for researchers both before and after their computational research is done. As with HPC interfaces, another requirement is that data storage and retrieval is user-friendly.

Data Oasis

Among Academia's Fastest Parallel File Systems

In 2012, SDSC "supercharged" *Data Oasis*, a Lustre-based parallel file storage system linked to *Trestles*, *TSCC*, and *Gordon*. As a critical component of SDSC's Big Data initiatives, *Data Oasis* currently has four petabytes of capacity and speeds of up to 100 gigabytes per second to handle just about any data-intensive project. Using the I/O power of *Gordon* and *Trestles*, those sustained transfer rates make *Data Oasis* one the fastest parallel file systems in the academic community. The sustained speeds mean researchers could retrieve or store 64 terabytes (TB) of data—the equivalent of *Gordon's* entire DRAM memory—in about 10 minutes, significantly reducing time needed for retrieving, analyzing, storing, or sharing extremely large datasets.

Big Data is not just about sheer size, but also about the speed of moving data where it needs to be, and the integrated infrastructure and software tools needed to effectively do research using those data. The capability of *Data Oasis* allows researchers to analyze data at a much faster rate than other systems, which in turn helps extract knowledge and discovery from these datasets.

SDSC CLOUD

Version 2.0

The *SDSC Cloud*, which went into production in late 2011, is the first large-scale academic deployment of cloud storage in the world. *SDSC Cloud* storage currently has 2.7 petabytes of raw space used by almost 400 partners. The system is being expanded from an object-based file store running on OpenStack's SWIFT platform to include other capabilities. *SDSC Cloud 2.0* includes Nova, OpenStack's cloud compute capability. This provides the opportunity to build Data as a Service (DaaS) on top of an open source, cloud-based technology, marrying SDSC's service offerings with its growing expertise in cloud computing and solid data management and engineering research. The platform was designed around the use cases of researchers, which includes SDSC's PIs working on data management and Big Data projects as well as campus, system, and collaborators. In 2014, SDSC plans to launch the PACE Starter Kit, a collection of cloud-based and virtual technologies to extend data mining sandboxes and environments following the popular Data Mining Boot Camps held by SDSC's Predictive Analytics Center of Excellence (PACE).

Christine Kirkpatrick is division director of SDSC's IT Systems and Services Division, which designs, deploys, and operates high-performance systems and provides services supporting a full range of academic and industry researchers.

NETWORKING/CONNECTIVITY

While rapid computational processing and stable storage structures are essential to conduct data-enabled science, the cyberinfrastructure must also include networks that allow fast and unrestricted flow of information between systems and researchers. The answer is using dedicated 'fat' pipes that can accommodate extreme-sized bursts of data. SDSC has helped lead the way on several fronts in this area.

Prism@UCSD

The HOV Lane for Broad-bandwidth Research

Working with campus partners, SDSC helped establish a research-defined, end-to-end networking cyberinfrastructure for the UC San Diego campus that is capable of supporting large data transmissions between facilities that might otherwise hobble the main campus network. Called Prism@UCSD and backed by a $500,000 NSF grant, researchers with the campus' California Institute for Telecommunications and Information Technology (Calit2) and SDSC began work on the network in 2013 to support research in data-intensive areas such as genomic sequencing, climate science, electron microscopy, oceanography, and physics. Slated for completion in late 2014, the project is already serving researchers with full functionality.

"One can think of Prism as the HOV lane, whereas our very capable campus network represents the other lanes on the freeway," says Philip Papadopoulos, principal investigator on the Prism@ UCSD project and SDSC's chief technology officer.

Almost 150 communication fibers tunnel into UC San Diego's Atkinson Hall, including the main Calit2 server room (pictured above). From there, researchers have access to the main campus networks, as well as regional, national, and international wireless and optical networking test beds and visualization centers.

CHERuB

Connecting to the Information Superhighway

In late 2013 SDSC and UC San Diego's Administrative Computing and Telecommunications (ACT) organization were awarded a second $500,000 NSF grant to connect the campus to high-bandwidth national research networks to advance a new range of data-driven research. Called CHERuB for Configurable, High-speed, Extensible Research Bandwidth, the project will provide 100 gigabit-per-second connectivity—the new high-end for wide-area research networks. CHERuB will support multi-institutional data transit over networks such as the Internet2's Advanced Layer 2 Service (AL2S) and ESnet, as well as a joint project between those networks called the Advanced Networking Initiative (ANI), the result of a $62 million grant under the American Recovery and Reinvestment Act to build a national 100G "information backbone."

When completed, the CHERuB link will place UC San Diego among research universities and institutions having the highest available connectivity, with a capacity 10 times greater than existing modern data networks. CHERuB is the missing piece that will connect UC San Diego's Prism network to even faster national networks to advance scientific research.

Examples of research domains that will benefit from CHERuB include cosmology, atmospheric sciences, electron microscopy, genomic sequencing, oceanography, high-energy physics, and telemedicine—all of which can encompass data-rich research and whose advancements rely on multi-site or inter-institutional activities.

SCIENCE HIGHLIGHTS

TARGETING DISEASE *with* “Designer” Drugs

Finding and understanding how proteins bind to one another to initiate disease processes has been one of the most critical tasks performed in recent years by high-performance computing and algorithms to speed those discoveries. The work is fundamental to drug discovery and what has become commonly referred to as “drug design.”

Igor Tsigelny, a research scientist at SDSC, has been using SDSC computational resources to help experts search for new insights into a cross-section of medically challenging conditions, from heart disease to cancer to mental disorders.

During the past year, for example, Tsigelny worked with other researchers at UC San Diego and the Institut Pasteur in Paris to identify coherent-gene-groups (CGGs) responsible for brain development which can be affected for the treatment of developmental and mental disorders such as autism-spectrum disorders (ASD) and schizophrenia. In a paper published in the journal *Gene, Brain and Behavior*, the researchers identified the hierarchical tree of CGG-transcription factor (TF) networks that determine the patterns of genes expressed during brain development, and found that some "master transcription factors" at the top of the hierarchy regulated the expression of a significant number of gene groups.

Using samples taken from three different regions of the brains of rats, the researchers used *Gordon* and SDSC's BiologicalNetworks server to conduct numerous levels of analyses, starting with processing of microarray and SOM (self-organizing maps) clustering, before determining which gene zones were associated with significant developmental changes and brain disorders.

Said Tsigelny, also a researcher with UC San Diego Moores Cancer Center and the Department of Neurosciences: "We have proposed a novel, though still hypothetical, strategy of drug design based on this hierarchical network of TFs that could pave the way for a new category of pharmacological agents that could be used to block a pathway at a critical time during brain development as an effective way to treat and prevent mental disorders such as ASD and schizophrenia. On a broader scale, these findings have the potential to change the paradigm of drug discovery."

Tsigelny's collaborators included Valentina L. Kouznetsova (SDSC and Moores); Michael Baitaluk (SDSC); and Jean-Pierre Changeux, with the Institut Pasteur, a distinguished visiting professor in pharmacology at UC San Diego (2008) and member of the foreign faculty at UC San Diego's Kavli Institute for Brain and Mind.

In 2013, Tsigelny was a part of a team of UC San Diego researchers that designed new compounds that mimic those naturally used by the body to regulate blood pressure. The most promising of them may literally be the key to controlling hypertension, switching off the signaling pathways that lead to this condition.

Published in the online version of *Bioorganic & Medicinal Chemistry*, the scientists studied the properties of the peptide called catestatin that binds nicotinic acetylcholine receptors found in the nervous system, and developed a pharmacophore model of its active centers. They next screened a library of compounds for molecules that might match this 3D "fingerprint". The scientists then took their *in-silico* findings and applied them to lab experiments, uncovering compounds that successfully lowered hypertension.

The research may lead to a new class of treatments for hypertension, a disease which affects about 76 million people, or about one in three adults, in the United States, according to the American Heart Association.

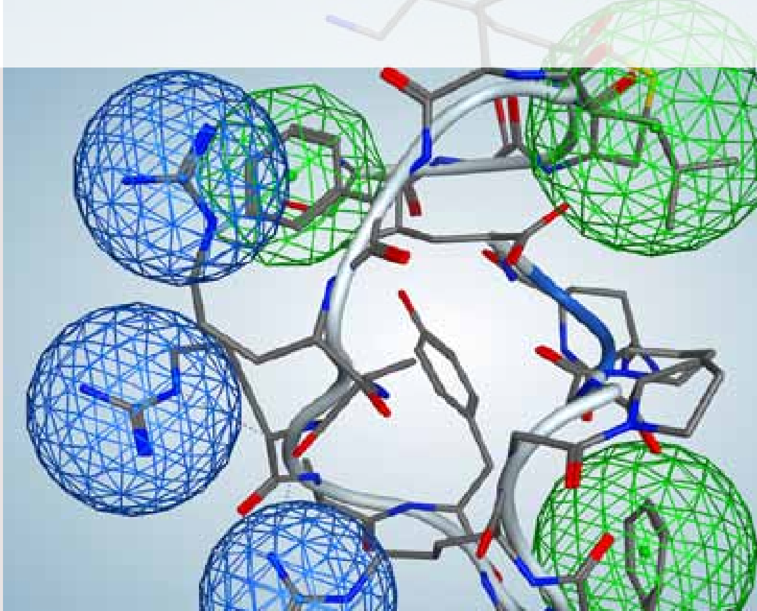
"Our results suggest that analogs can be designed to match the action of catestatin, which the body uses to regulate blood pressure," said Daniel T. O'Connor, a professor at the UC San Diego School of Medicine and senior author of the study.

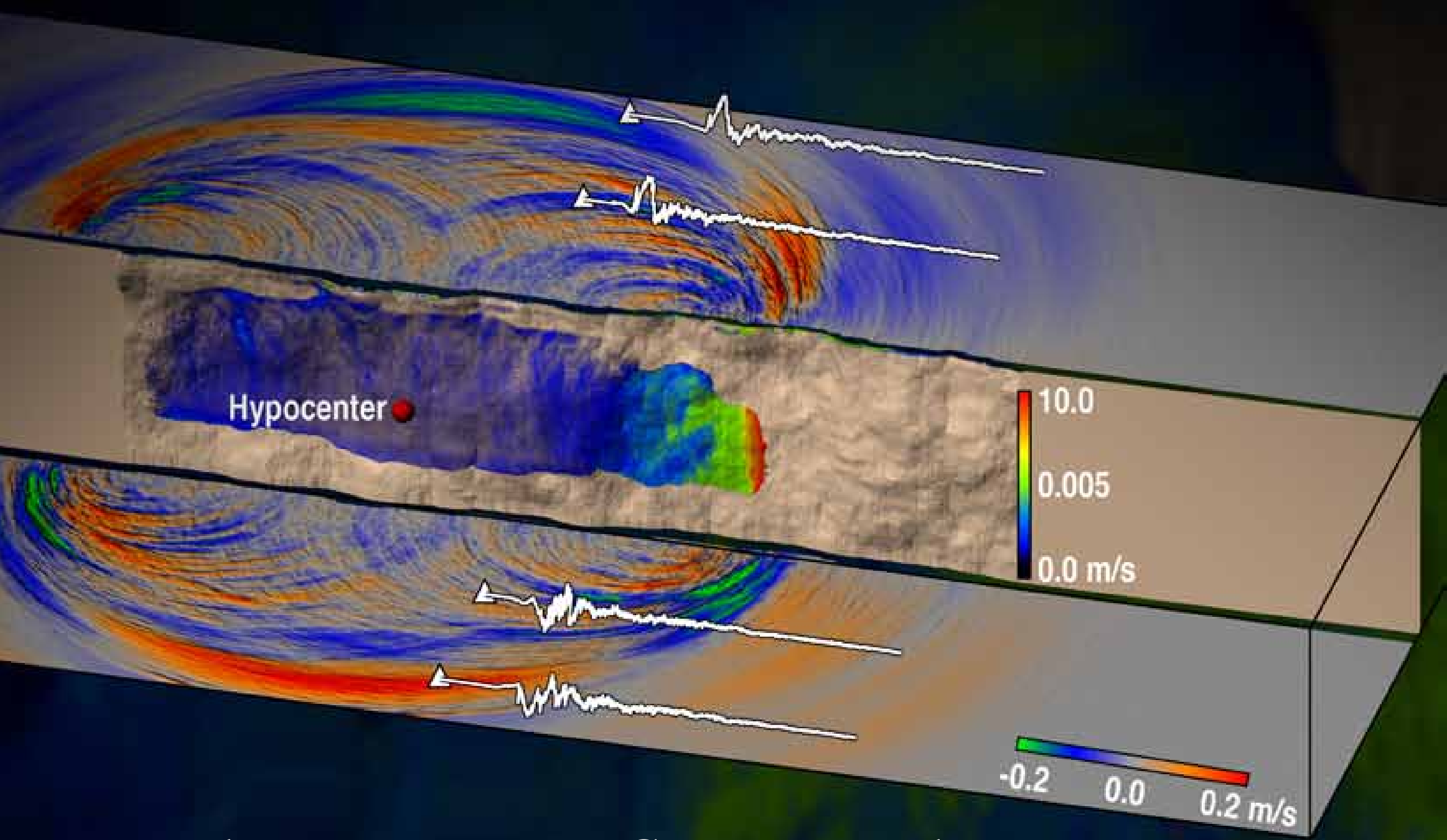
"This approach demonstrates the effectiveness of rational design of novel drug candidates," added Tsigelny.

Other authors included Kouznetsova, Nilima Biswas and Sushil K. Mahata, of UC San Diego's Departments of Medicine and Pharmacology.

Igor Tsigelny, SDSC research scientist (above)

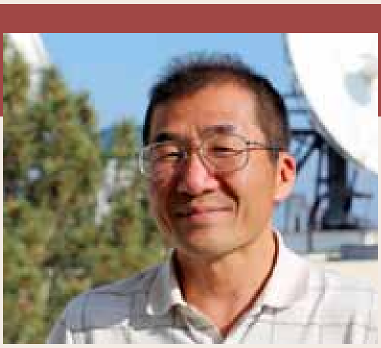
Catestatin-mimic pharmacophore model developed by researchers to help in the fight against hypertension. Image: Valentina Kouznetsova, UC San Diego (right)





Achieving Petaflop-Level EARTHQUAKE SIMULATIONS on GPU-Powered Supercomputers

Image of a 10-Hz rupture propagation and surface wavefield for a crustal model with a statistical model of small-scale heterogeneities. Simulation by Yifeng Cui and Efecan Poyraz; visualization by Amit Chourasia, SDSC.



Yifeng Cui (left) is an SDSC computational scientist specializing in high-performance earthquake simulations, as well as parallelization, optimization, and performance evaluation on both massively parallel and vector machines.

Dong Ju Choi (right) is a senior computational scientist with diverse expertise in high-performance computing software, programming, optimization, and visualization.

To save lives and minimize property damage resulting from earthquakes, researchers are seeking to rapidly assimilate vast quantities of information from earthquake cascades to improve operational forecasting and provide early warning systems.

A step toward that goal was achieved last year by a team of researchers at SDSC and the Department of Electronic and Computer Engineering at UC San Diego with the development of a highly scalable computer code that promises to dramatically cut both research times and energy costs needed to simulate seismic hazards throughout California and elsewhere.

The team, led by Yifeng Cui, a computational scientist at SDSC, developed the scalable GPU (graphical processing units) accelerated code for use in earthquake engineering and disaster management through regional earthquake simulations at the petascale level as part of a larger computational effort coordinated by the Southern California Earthquake Center (SCEC). San Diego State University (SDSU) is also part of this collaborative effort in pushing toward extreme-scale earthquake computing.

"The increased capability of GPUs, combined with the high-level GPU programming language CUDA, has provided tremendous horsepower required for acceleration of numerically intensive 3D simulation of earthquake ground motions," said Cui, who presented the team's new development at the NVIDIA 2013 GPU Technology Conference (GTC) in San Jose, Calif. A technical paper based on this work was also presented June 5-7 at the 2013 International Conference on Computational Science in Barcelona, Spain.

The accelerated code, done using GPUs as opposed to CPUs, or central processing units, is based on a widely-used wave propagation code called AWP-ODC, which stands for Anelastic Wave Propagation by Olsen, Day, and Cui. It was named after Kim Olsen and Steven Day, geological science professors at SDSU, and SDSC's Cui. The research team restructured the code to exploit high performance and throughput, memory locality, and overlapping of computation and communication, which made it possible to scale the code linearly to more than 8,000 NVIDIA Kepler GPU accelerators.

The team performed GPU-based benchmark simulations of the 5.4 magnitude earthquake that occurred in July 2008 below Chino Hills, near Los Angeles. Compute systems included *Keeneland*, managed by Georgia Tech, Oak Ridge National Laboratory (ORNL) and the National Institute for Computational Sciences (NICS), and also part of the National Science Foundation's (NSF) eXtreme Science and Engineering Discovery Environment (XSEDE); and *Blue Waters*, based at the National Center for Supercomputing Applications (NCSA). Also used was the *Titan* supercomputer, based at ORNL and funded by the U.S. Department of Energy. *Titan* is equipped with Cray XK7 systems and NIVIDIA's Tesla K20X GPU accelerators.

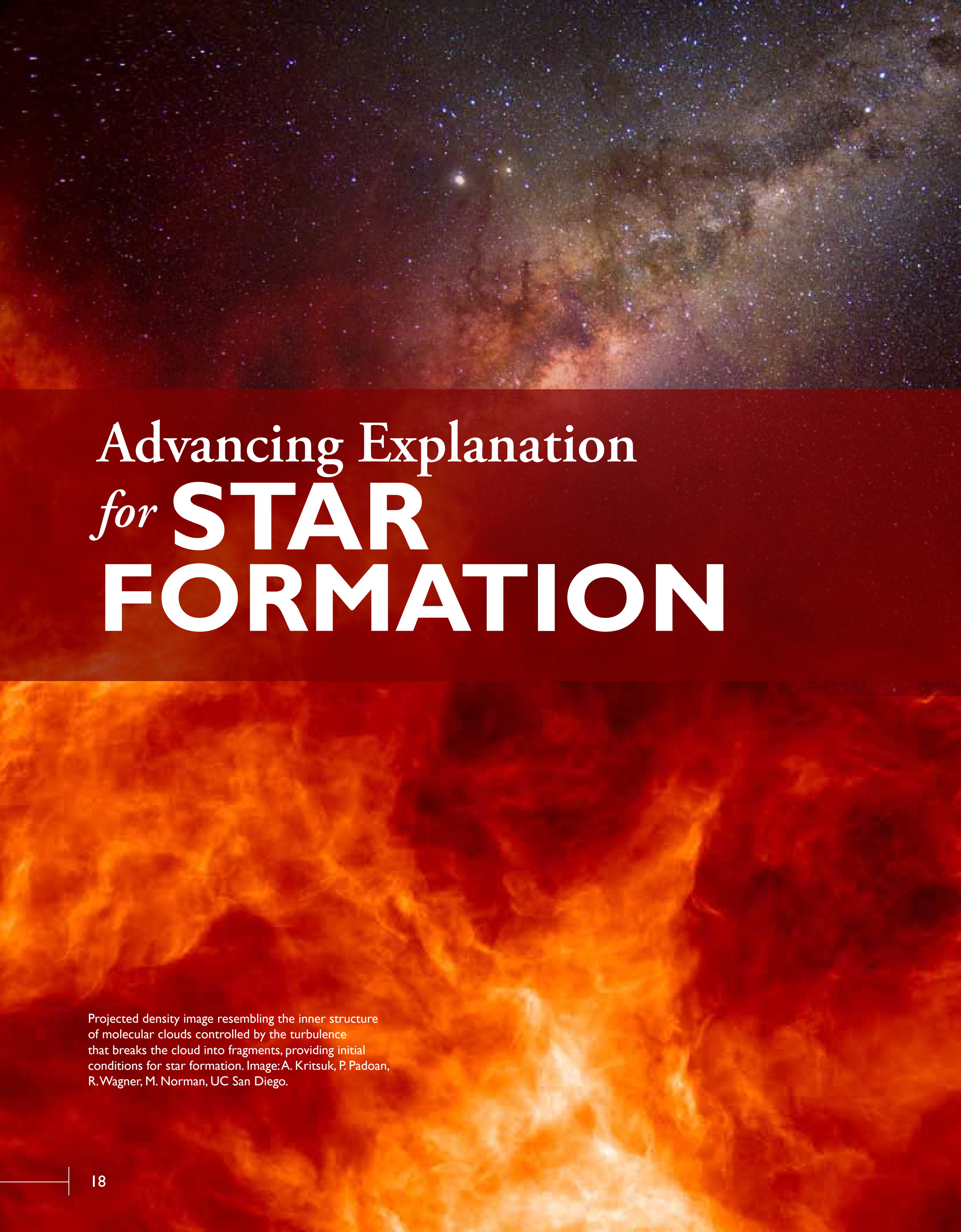
The benchmarks, run on *Titan*, showed a five-fold speedup over the heavily optimized CPU code on the same system, and a sustained performance of one petaflop per second (one quadrillion calculations per second) on the tested system. A previous benchmark of the AWP-ODC code reached only 200 teraflops (trillions of calculations per second) of sustained performance.

By delivering a significantly higher level of computational power, researchers can provide more accurate earthquake predictions with increased physical reality and resolution, with the potential of saving lives and minimizing property damage.

"This is an impressive achievement that has made petascale-level computing a reality for us, opening up some new and really interesting possibilities for earthquake research," said Thomas Jordan, director of SCEC, which has been collaborating with UC San Diego and SDSU researchers on this and other seismic research projects, such as the simulation of a magnitude 8.0 earthquake, the largest ever simulation to-date.

Additional members on the UC San Diego research team include Jun Zhou and Efecan Poyraz, graduate students with the university's Department of Electrical and Computer Engineering (Zhou devoted his graduate research to this development work); SDSC researcher Dong Ju Choi; and Clark C. Guest, an associate professor of electrical and computer engineering at UC San Diego's Jacobs School of Engineering.

Advancing Explanation *for* STAR FORMATION



Projected density image resembling the inner structure of molecular clouds controlled by the turbulence that breaks the cloud into fragments, providing initial conditions for star formation. Image: A. Kritsuk, P. Padoan, R. Wagner, M. Norman, UC San Diego.

Alexei Kritsuk is a research physicist with UC San Diego's Physics Department and Center for Astrophysics & Space Sciences (CASS)

Insights gleaned from six supercomputer simulations of interstellar medium, at SDSC and elsewhere, last year confirmed principles in a seminal paper published in 1981 describing essential relationships of structure and motion in molecular clouds where stars form.

The new analysis, described in the October 2013 issue of the *Monthly Notices of the Royal Astronomical Society*—Great Britain's pre-eminent astronomy and astrophysics journal—provided for the first time an explanation for the origin of three observed correlations between various properties of molecular clouds in the Milky Way known as Larson's Laws, named for Richard Larson, now an Emeritus Professor of Astronomy at Yale University.

"After decades of inconclusive debate about the interpretation of the correlations among molecular cloud properties that I published in 1981, it's gratifying to see that my original idea that they reflect a hierarchy of supersonic turbulent motions is well supported by these detailed new simulations," said Larson in response to the new findings by three UC San Diego astrophysics researchers.

"This paper is essentially the culmination of seven years of research, aided by the use of large-scale supercomputer simulations conducted at SDSC and elsewhere," said Alexei Kritsuk, a research physicist with UC San Diego's Physics Department and Center for Astrophysics & Space Sciences (CASS), and lead author of the paper. "Molecular clouds are the birth sites for stars, so this paper relates also to the theory of star formation."

The analysis by the UC San Diego researchers is based on recent observational measurements and data from six simulations of the interstellar medium, including the effects of self-gravity, turbulence, magnetic field, and multiphase theramodynamics. The supercomputer simulations support a turbulent interpretation of Larson's relations, and the study concludes that there are not three independent Larson laws, but that all three correlations are due to the same underlying physics, (i.e. properties of supersonic turbulence).

Larson's original paper, published in the same journal, still inspires new advances in the understanding of molecular cloud structure formation and star formation.

The research team included Michael Norman, SDSC Director and a Distinguished Professor of Physics at UC San Diego; and Christoph T. Lee, an undergraduate researcher with CASS. SDSC's *Trestles* and *Triton* clusters, and now-decommissioned *DataStar* systems, were used to generate the simulations, as well as the *Kraken* and *Nautilus* systems at the National Institute for Computational Science (NICS) at Oak Ridge National Laboratory.

"None of these new findings and insights would have been possible without the tremendous advances in supercomputer simulations that allow not only cosmologists but scientists in countless other domains an unprecedented level of resolution and data-processing speed to further their research," said Norman, a globally recognized astrophysicist who has pioneered the use of advanced computational methods to explore the universe and its beginnings.

"We believe that this paper paints the complete picture, drawing from earlier published works of ours as well as presenting new simulations that have not been published before," Norman added.

Small Magellanic Cloud. Image credit NASA/CXC/JPL-Caltech/STScl.

Crunching Large Hadron COLLIDER DATA to Speed Dark Matter Quest

With the discovery and later confirmation last March of the Higgs boson—the last missing piece of the standard model of particle physics—scientists are now setting their sights on discovering new physics beyond the standard model. The next big thing is to search for dark matter, according to Frank Wuerthwein, a professor of physics at UC San Diego.

"For the Higgs, we knew exactly how to search for it given theoretical predictions based on past experimental results," said Wuerthwein, part of the research team for the Compact Muon Solenoid (CMS), one of two large general-purpose particle detectors at the Large Hadron Collider (LHC) used by researchers in Switzerland to find the elusive Higgs boson.

"For dark matter, the situation is much hazier," Wuerthwein added. "We hope to produce dark matter at the LHC in cascade decays of a whole spectrum of new fundamental particles, the lowest mass of which is dark matter. But the details of this spectrum of masses are unknown. To have sensitivity to a large range of possible mass spectra, we needed to write more data to tape so we would be able to carefully analyze it later."

Toward that end, a team of UC San Diego physicists including Wuerthwein, and the Open Science Grid (OSG), a multi-disciplinary research partnership funded by the U.S. Department of Energy and the National Science Foundation, used SDSC's *Gordon* to rapidly process raw data from almost one billion particle collisions generated by the CMS. The project represented the single most data-intensive exercise to date for *Gordon* since completing its large-scale acceptance testing in early 2012.

The UC San Diego-OSG collaboration (see page 41) came as the LHC was shut down in February 2013 to make numerous upgrades during the next two years. One major activity during the shutdown included the development of plans for efficient, effective searches once the LHC was back in operation. To do that—and to have time enough to upgrade equipment—researchers had to sift through massive amounts of stockpiled data to help scientists define future research agendas, such as the search for dark matter.

"Access to *Gordon*, and its excellent computing speed due to its flash-based memory, really helped push forward the processing schedule for us," said Wuerthwein. "With only a few weeks' notice, we were able to gain access to *Gordon* and complete the runs, making the data available for analysis in time to provide crucial input toward international planning meetings on the future of particle physics."

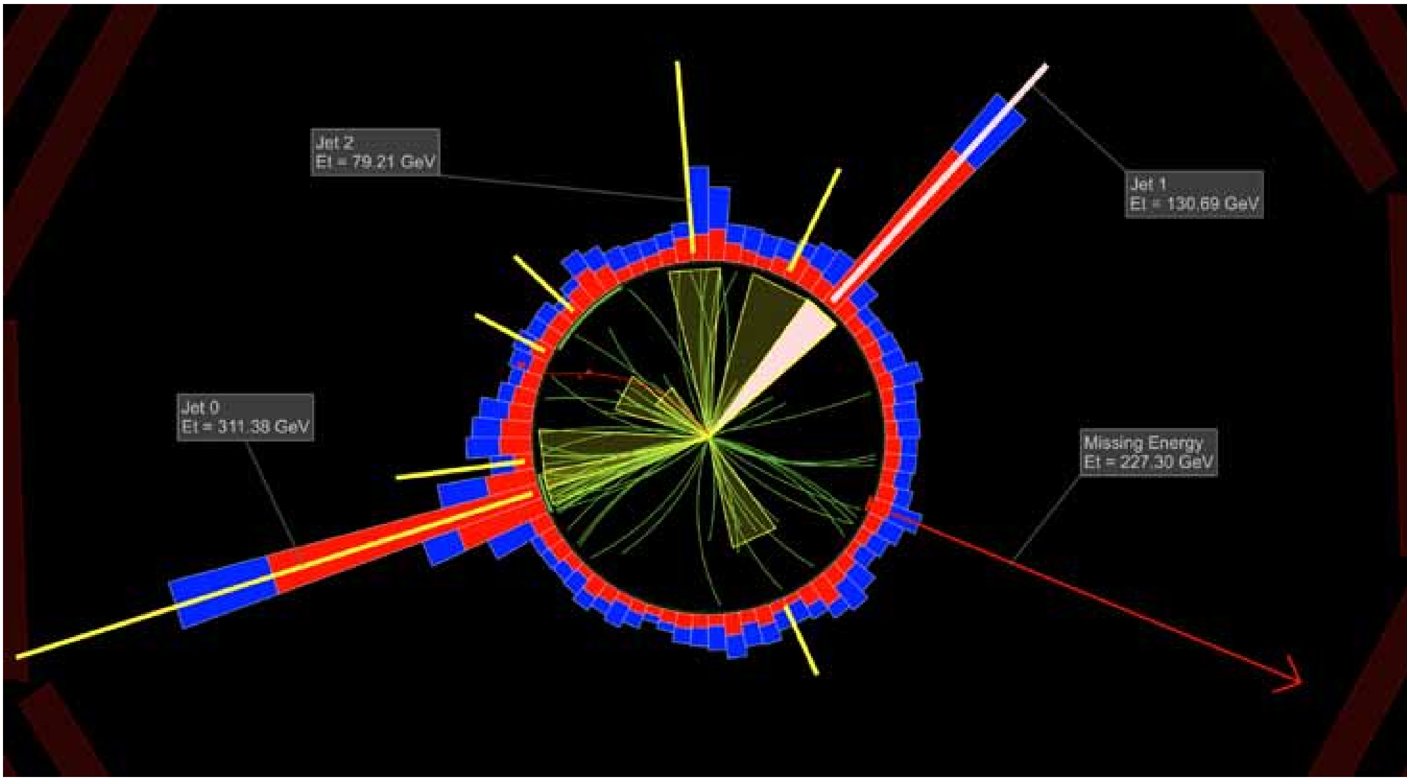
"Giving us access to the *Gordon* supercomputer effectively doubled the data processing compute power available to us," added Lothar Bauerdick, OSG's executive director and the U.S. software and computing manager for the CMS project. "This gave CMS scientists precious months to get to their science analysis of the data reconstructed at SDSC."

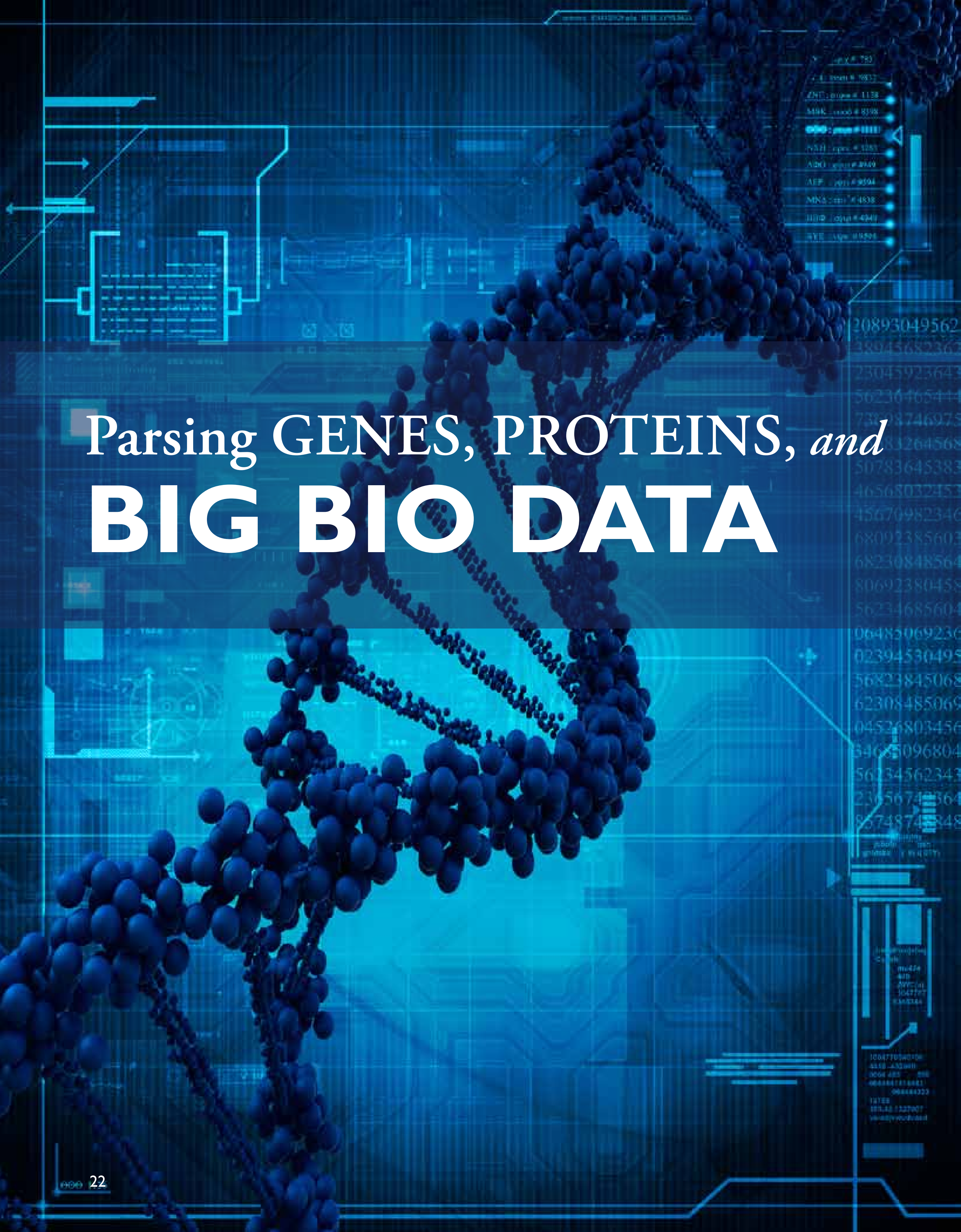
UC San Diego researchers and CMS team members, in addition to Wuerthwein, included Jim Branson, Vivek Sharma, and Avi Yagil, all of whom played major roles in the discovery of the Higgs boson particle.

"UC San Diego has been one of the most successful institutions in the global hunt for the Higgs particle discovery at the LHC," said Wuerthwein, who is leading the university's search for dark matter.

Amit Majumdar is interim director of SDSC's Data Enabled Scientific Computing (DESC) division, which assisted researchers in using *Gordon* to process data on almost one billion particle collisions.

This image of a supersymmetry event shows the transverse momentum imbalance due to dark matter particles escaping the detector (direction indicated by red arrow). Red and blue rectangles indicate energy deposited in the electromagnetic and hadronic calorimeter respectively; green tracks in the center show charged particles with transverse momentum larger than 2GeV. Yellow-outlined triangles indicate jet cones or the presence of subatomic particles called quarks. Image courtesy of Matevz Tadel, UC San Diego/CMS





Parsing GENES, PROTEINS, *and* BIG BIO DATA

Wayne Pfeiffer is an SDSC Distinguished Scientist. His current research focuses on computationally challenging problems in bioinformatics, including genomic variation analyses.

SDSC's data-intensive computing resources have proven to be a boon to biologists interested in rapidly sifting through ever-expanding amounts of data or trying to tame the tidal wave of genomic data used to sequence the DNA of an organism, whether human, plant, or jellyfish.

"Next-generation sequencing has profoundly transformed biology and medicine, providing insight into our origins and diseases," according to Wayne Pfeiffer, a Distinguished Scientist at SDSC. "However, obtaining that insight from the data deluge requires complex software and increasingly powerful computers."

Available for use by industry and government agencies, SDSC's *Gordon* and *Trestles* are part of the NSF's (National Science Foundation) XSEDE (eXtreme Science and Engineering Discovery Environment) program, a nationwide partnership comprising 16 supercomputers as well as high-end visualization and data analysis resources.

Following are examples of how *Gordon* and *Trestles*, as well as the development of "science gateways" that give researchers Web-based access to these and other HPC systems, are improving the scope and accessibility of essential biological research databases, while creating faster and more effective ways to assemble genomic information.

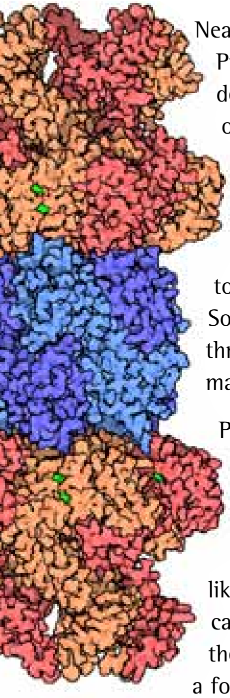
One Search to Bind Them: IntegromeDB

The diversity of biological fields has spawned thousands of databases and millions of public biomedical, biochemical, and drug and disease-related resources. For researchers interested in collecting information from those resources, search engines such as Google are of limited use since they are unable to comprehend the language of biology. They return results on the basis of keywords rather than in terms of scientific importance.

Michael Baitaluk and Julia Ponomarenko, principal investigators at SDSC, created a "smart" search for biologists, one able to return gene- and protein-centered knowledge in a biologically meaningful way—for example, by pathways, binding partners, structures, mutations, associated diseases, splice variants, or experiments. Called IntegromeDB for its ability to integrate biomedical data, the resource includes more than 16 million experimental findings. Receptor binding data for drugs and bioactive compounds, kinetic information for drug-metabolizing enzymes, and relevant signaling proteins are semantically linked to nearly 120 ontologies with a controlled vocabulary of approximately 70 million synonyms. Since its launch in January 2012, more than 4,000 users have visited the resource, and it has become an official science gateway for the NSF.

Stored in a PostgreSQL database, IntegromeDB contains more than 5,000 tables, 500 billion rows, and 50 terabytes of data. Baitaluk predicts IntegromeDB will eventually require more than a single petabyte of storage. The resource utilizes 16 compute and four I/O nodes of *Gordon* and 150 terabytes on SDSC's *Data Oasis* storage system.

Unclogging a Bottleneck for the Protein Data Bank



Nearly 250,000 scientists take advantage of the RCSB Protein Data Bank each month, every one of them depending on the resource to quickly provide details on more than 90,000 proteins, nucleic acids, and complex assemblies. While the PDB's current resources can easily handle research requests such as pairwise protein comparisons, the calculation of large numbers of protein structure alignments is too computationally intensive to be done in real time. So the PDB pre-calculates a large number of pairwise three-dimensional protein structure alignments and makes them available via its website.

Periodically, those alignments are recalculated as new protein structures and deposited into the database. However, the process of updating slows at the server that stands between the PDB and the nodes used to perform the alignment calculations. Much like an overwhelmed traffic interchange, the system cannot keep up with the data going and coming from the database, creating a data traffic jam. Phil Bourne, a former professor of pharmacology at UC San Diego, and Andreas Prlić, a senior scientist with the university, investigated whether this process could be improved using *Gordon*.

They found that the calculations were sped up 3.8 times; previous calculations that required 24 hours now took 6.3 hours. To gauge whether I/O performance might improve even further, Bourne and his colleagues tried the same calculations using Intel's new Taylorsville flash drives. The drives delivered twice as much bandwidth and read IOPS, and 13 times more write IOPS than Intel's Lyndonville flash drives. This drove the time down to 4.1 hours.

"With its excellent communications capabilities, *Gordon* can be used to greatly reduce the time to solution over the systems we currently use," said Bourne.

Taming the tidal wave of genomic data

Knowing the whole genome of various species underlies biological and medical research, such as understanding evolution pathways or identifying the causes of diseases. However, existing sequencing techniques produce huge amounts—billions for a high organism such as a human—of overlapping short sequences randomly sampled from the genome. A major challenge in genome research is to assemble these short reads, which vary from ten to several hundred bases, back into the whole genome, a task that requires vast amounts of memory. It would be similar to gluing together an encyclopedia from a haystack of words and sentence fragments.

Using *Trestles*, Xifeng Yan, the Venkatesh Narayanamurti Chair of Computer Science at the University of California, Santa Barbara, and his colleagues demonstrated that a new algorithm called MSP reduces one of the steps required so that it uses significantly less memory, a mere 10 gigabytes, than widely used algorithms. The results promise to remove one of the bottlenecks to processing whole genomes, thus making it possible to assemble large genomes using smaller, less expensive, commodity clusters.

"High-quality genome sequencing is foundational to many critical biological and medical problems," said Yan. "With the advent of massively parallel DNA sequencing technologies, how to manage and process the big sequence data has become an important issue. Experimental results showed that MSP can not only successfully complete the tasks on very large datasets within a small amount of memory, but also achieve better performance than existing state-of-the-art algorithms."

AAA+ Protease, image courtesy of the RCSB Protein Data Bank (above)

SDSC's *Trestles* supercomputer (right)

Mark Miller leads the Next Generation Tools for Biology group at SDSC and is founder of the CIPRES science gateway.

Reducing Research Barriers through Science Gateways

Making supercomputers more accessible to researchers is another area of focus at SDSC. One solution is the development of science gateways, or virtual environments that provide researchers with web-based access to tools, applications, computing resources, and data archives to further their scientific studies. Researchers can access top-tier resources, such as applications running on a supercomputer, remote instruments such as telescopes or electron microscopes, or curated data collections.

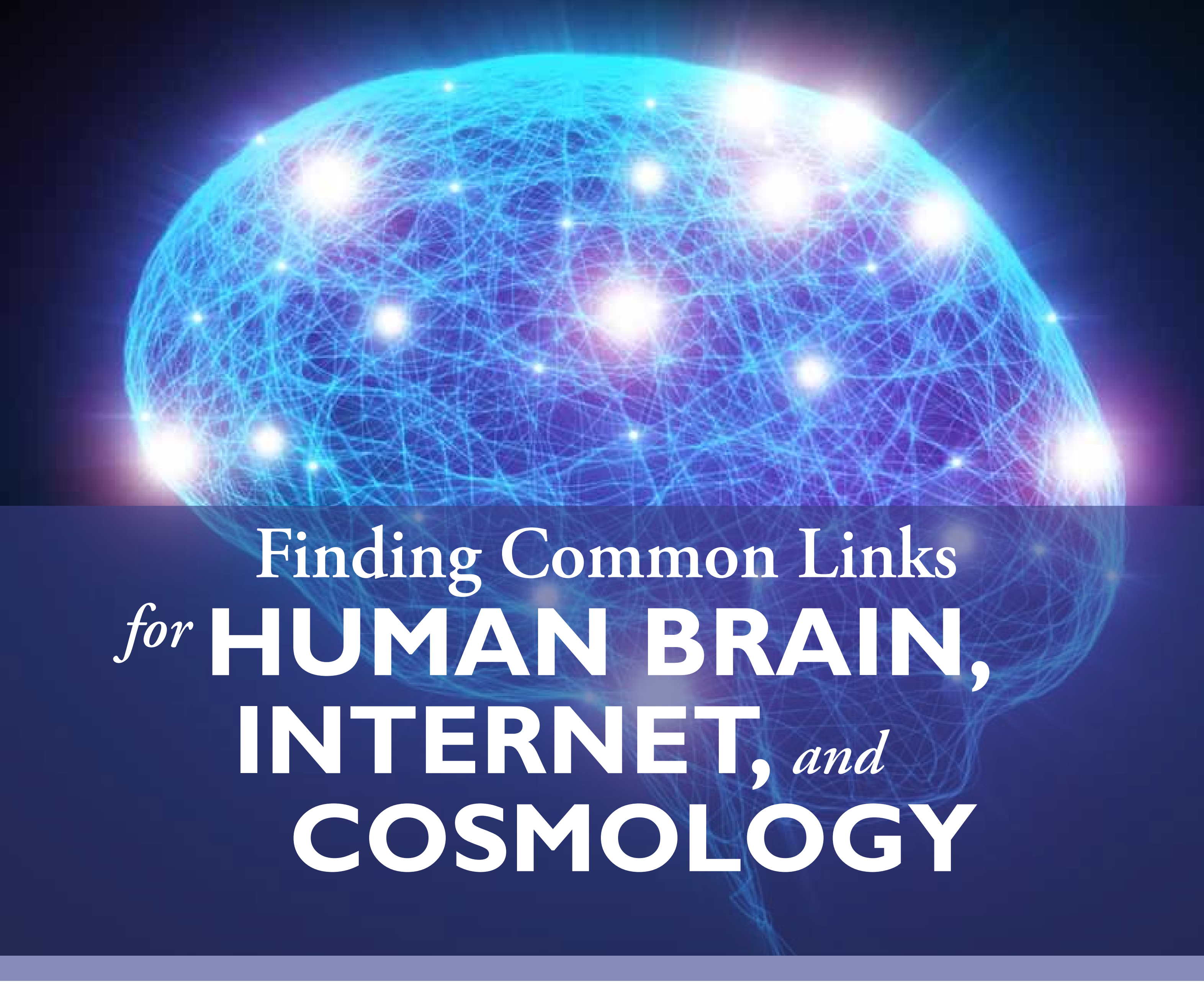
SDSC last year received a $1.5 million NSF award to make access to supercomputing resources simpler and more flexible for phylogenetic researchers. The award, which follows an earlier NSF grant that ran from 2003 to 2008, was for the CIPRES Science Gateway, a web site that allows researchers to explore evolutionary relationships between species using SDSC supercomputers as well as systems in XSEDE's repertoire. CIPRES stands for CyberInfrastructure for Phylogenetic Research.

"The CIPRES Gateway allows scientists to conduct their research in significantly shorter times without having to understand how to operate supercomputers," said Mark Miller, principal investigator for the gateway and an SDSC researcher. At of the end of 2013, the CIPRES Science Gateway supported more than 8,600 users and led to more than 700 publications of phylogenetic studies involving species in every branch of the Tree of Life.

Beyond phylogenetics, SDSC is a partner under a $5 million NSF grant to help build a science gateway service platform that will give researchers improved access to a variety of hosted or cloud services. Called SciGaP, the project is a collaboration among researchers at Indiana University and the University of Texas aimed at significantly lowering the development overhead for communities that wish to create new science gateways, allowing gateway creators to focus on developing new capabilities that are unique to an individual gateway's scientific community.

Science Gateway group leaders Miller and Amit Majumdar are leading SDSC's participation in the project. "With the SciGaP project we hope to enable a large number of existing and new science gateways from various domain sciences," said Majumdar, interim director of SDSC's Data Enabled Scientific Computing division.

Tree of Life image courtesy of Nick Kurzenko, Greg Rouse, and the U. S. Fish and Wildlife Service.



Finding Common Links *for* HUMAN BRAIN, INTERNET, *and* COSMOLOGY

The structure of the universe and the laws that govern its growth may be more similar than previously thought to the structure and growth of the human brain and other complex networks, such as the Internet or a social network of trust relationships between people, according to a paper published last year in the science journal *Nature's Scientific Reports*.

"By no means do we claim that the universe is a global brain or a computer," said Dmitri Krioukov, co-author of the paper, published by the Cooperative Association for Internet Data Analysis (CAIDA), based SDSC. "But the discovered equivalence between the growth of the universe and complex networks strongly suggests that unexpectedly similar laws govern the dynamics of these very different complex systems."

Having the ability to predict—let alone trying to control —the dynamics of complex networks remains a central challenge throughout network science. Structural and dynamical similarities among different real networks suggest that some universal laws might be in action, although the nature and common origin of such laws remain elusive.

By performing complex supercomputer simulations of the universe and using a variety of other calculations, researchers have now proved that the causal network representing the large-scale structure of space and time in our accelerating universe is a graph that shows remarkable similarity to many complex networks such as the Internet, social, or even biological networks.

"These findings have key implications for both network science and cosmology," noted Krioukov. "We discovered that the large-scale growth dynamics of complex networks and causal networks are asymptotically (at large times) the same, explaining the structural similarity between these networks."

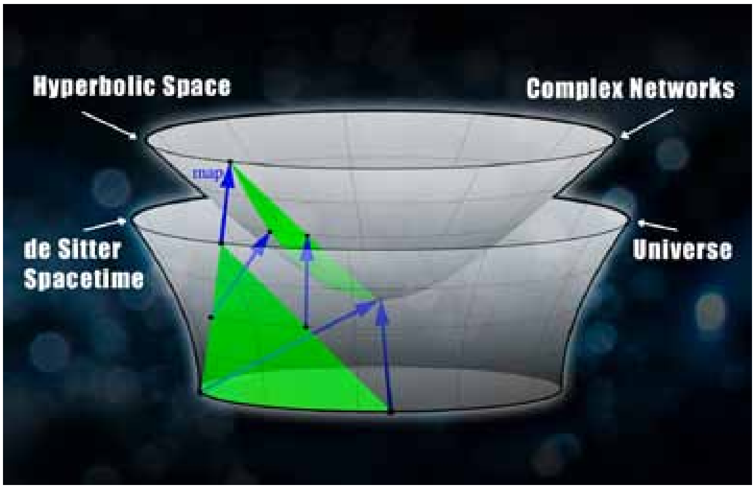
"This is a perfect example of interdisciplinary research combining math, physics, and computer science in totally unexpected ways," said SDSC Director Michael Norman. "Who would have guessed that the emergence of our universe's four-dimensional spacetime from the quantum vacuum would have anything to do with the growth of the Internet? Causality is at the heart of both, so perhaps the similarity Krioukov and his collaborators found is to be expected."

Of course the network representing the structure of the universe is astronomically huge—in fact it can be infinite. But even if it is finite, researchers' best guess is that it is no smaller than 10^{250} atoms of space and time. (That's the digit 1 followed by 250 zeros.) For comparison, the number of water molecules in all the oceans of the world has been estimated to be 4.4 x 10^{46}.

Yet the researchers found a way to downscale this humongous network while preserving its vital properties, by proving mathematically that these properties do not depend on the network size in a certain range of parameters, such as the curvature and age of our universe.

After the downscaling, the research team turned to *Trestles*, one of SDSC's data-intensive supercomputers, to perform simulations of the universe's growing causal network. By parallelizing and optimizing the application, Robert Sinkovits, director of SDSC's Scientific Applications Group, was able to complete in just over one day a computation that was originally projected to require three to four years.

"In addition to being able to complete these simulations much faster than previously ever imagined, the results perfectly matched the theoretical predictions of the researchers," said Sinkovits.



Simple mapping between the two surfaces representing the geometries of the universe and complex networks proves that their large-scale growth dynamics and structures are similar. Image courtesy of CAIDA/SDSC

The most frequent question that people may ask is whether the discovered asymptotic equivalence between complex networks and the universe could be a coincidence," said Krioukov. "Of course it could be, but the probability of such a coincidence is extremely low. Coincidences in physics are extremely rare, and almost never happen. There is always an explanation, which may be not immediately obvious."

"Such an explanation could one day lead to a discovery of common fundamental laws whose two different consequences or limiting regimes are the laws of gravity (Einstein's equations in general relativity) describing the dynamics of the universe, and some yet-unknown equations describing the dynamics of complex networks," added Marián Boguñá, a member of the research team from the Departament de Física Fonamental at the Universitat de Barcelona, Spain.

Other researchers who worked on this project are Maksim Kitsak, CAIDA/SDSC/UC San Diego; and David Rideout and David Meyer, Department of Mathematics at UC San Diego.

kc claffy is head of SDSC's CAIDA group, and has played a leading role in Internet research for more than a decade responding to industry, government, and academic needs in providing tools and analyses to promote a scalable global Internet infrastructure.

Fostering **Computations** on **NON-TRADITIONAL** RESEARCH

When SDSC's *Gordon* supercomputer debuted in early 2012, the system's architects envisioned that its innovative features—such as the first large-scale deployment of flash storage (300 terabytes) in a high-performance computer—would open the door to new areas of research.

"I view *Gordon* as a new kind of vessel, a ship that will take us on new voyages to make new discoveries in new areas of science," said Mike Norman, SDSC Director of *Gordon's* principal investigator, prior to its launch.

Fast forward to 2013, and it's clear that *Gordon* already reached out to areas relatively new to advanced computing, such as political science, mathematical anthropology, finance, and even the cinematic arts.

Following are three examples of what might be considered "non-traditional" computational research assisted by *Gordon* during the past year.

SDSC's innovative *Gordon* supercomputer made its debut in 2012. (above)

Gordon installation in the SDSC Datacenter (left)

Large Scale Video Analytics

Virginia Kuhn, associate director of the Institute for Multimedia Literacy and associate professor in the School of Cinematic Arts at the University of Southern California (USC), has been using *Gordon* to search, index, and tag vast media archives in real time, applying a hybrid process of machine analytics and crowd-sourced tagging.

The Large Scale Video Analytics (LSVA) project is a collaboration among cinema scholars, digital humanists, and computational scientists from the IML (Institute for Multimedia Literacy); ICHASS (Institute for Computing in the Humanities, Arts and Social Sciences); and NSF's (National Science Foundation)) XSEDE (eXtreme Science and Engineering Discovery Environment). The project is customizing the Medici content management system to apply various algorithms for image recognition and visualization into workflows that will allow real-time analysis of video.

"Contemporary culture is awash in moving images," said Kuhn, principal investigator for the research project. "There is more video uploaded to YouTube in a day than a single person can ever view in a lifetime. As such, one must ask what the implications are when it comes to the issues of identity, memory, history, or politics."

The LVSA project turned to *Gordon* for its extensive and easily accessed storage capacity, since video data collections can easily reach multiple terabytes in size. Work was initially performed using a dedicated *Gordon* I/O node, and later expanded to also include four dedicated compute nodes. "Persistent access to the flash storage in the I/O nodes has been critical for minimizing data access times, while allowing interactive analysis that was so important to this project," said Robert Sinkovits, director of SDSC's Scientific Applications Group.

The system enabled data subsets that were most heavily used to reside in areas that provided fast random access.

Predictive Analytics Using Distributed Computing

For operations on large data sets, the amount of time spent moving data between levels of a computer's storage and memory hierarchy often dwarfs the computation time. In such cases, it can be more efficient to move the software to the data rather than the traditional approach of moving data to the software.

Distributed computing frameworks such as Hadoop take advantage of this new paradigm. Data sets are divided into chunks that are then stored across the Hadoop cluster's data nodes using the Hadoop Distributed File System (HDFS), and the MapReduce engine is used to enable each worker node to process its own portion of the data set before the final results are aggregated by the master node. Applications are generally easier to develop in the MapReduce model, avoiding the need to directly manage parallelism using MPI, Pthreads, or other fairly low-level approaches.

Yoav Freund, a computer science and engineering professor at UC San Diego, specializes in machine learning, a relatively new field of research which bridges computer science and statistics. A recognized authority in Big Data analytics, Freund recently taught a graduate-level class in which students used a dedicated Hadoop cluster on *Gordon* to analyze data sets ranging in size from hundreds of megabytes to several terabytes. While student projects were diverse—analysis of temperature and airflow sensors on the UC San Diego campus, detection of failures in the Internet through the analysis of the communication between BGP (Border Gateway Protocol) nodes, and predicting medical costs associated with vehicle accidents—they had one need in common: the need for rapid turnaround of data analysis.

"Given the large data movement requirements and the need for very rapid turnaround, it would not have been feasible for the students to work through a standard batch queue," said Freund. "Having access to flash storage greatly reduced the time for random data access."

A *Gordon* I/O node and the corresponding 16 compute nodes were configured as a dedicated Hadoop cluster, with the HDFS mounted on the solid state drives (SSDs). To enable experimentation with Hadoop, SDSC also deployed MyHadoop, which allows users to temporarily create Hadoop instances through the regular batch scheduler.

Robert Sinkovits is the Director of SDSC's Scientific Applications Group and has responsibility for ensuring that data-intensive problems can make effective use of this innovative flash memory-based system.

Modeling Human Societies

Doug White, a professor of anthropology at UC Irvine, is interested in how societies, cultures, social roles, and historical agents of change evolve dynamically out of multiple networks of social action. These networks can be understood using graph theory, where individuals are represented as vertices and the relationships between them are represented using edges.

In the 1990s, White and his colleagues discovered that networks could be meaningfully analyzed by a new algorithm that derives from one of the basic theorems in graph theory: the Menger vertex-connectivity theorem announced in 1927.

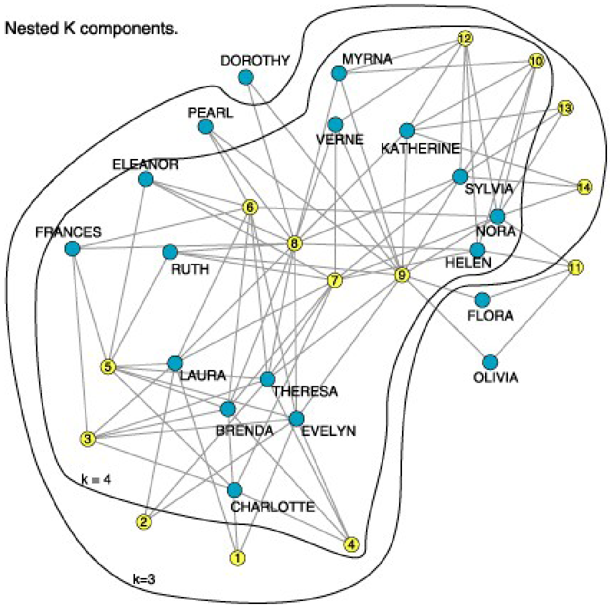
Computer scientists had previously coded Menger's edge-connectivity theorem, known as the Max-Flow (Ford-Fulkerson) network algorithm, to optimize the routing of traffic. But Menger's theorem also proved that vertex-connectivity, defined as the number of independent paths between a pair of vertices that share no vertices (other than the end points) in common, is equivalent to the number of vertices that need to be removed from the graph so that there is no longer a path between the two vertices.

Using Menger's theorem, it is possible to identify cohesive groups in which all members remain joined by paths as vertices are removed from the network. Finding the boundaries of cohesive subgroups, however, remain a challenge for computation, far more challenging than finding cliques of fully connected sets of vertices within a graph. White and his team are currently testing their recently developed software on synthetic networks constructed from well-known models, and plan to apply their methods to data sets derived from networks of co-authorship, benefits of marriage alliances, and splits in the scientific citation networks in periods of contention.

"The most computationally difficult step in the team's design of new high-performance software requires calculating the number of vertex-independent paths between all pairs of vertices in the network," said White.

Robert Sinkovits, director of SDSC's Scientific Applications Group, addressed scalability issues in the parallel algorithm and ported the application to *Gordon's* vSMP nodes following guidelines provided by ScaleMP. For larger graphs, such as a 2400 vertex network built using the Watts-Strogatz small-world model, near linear scalability was achieved when using all 256 cores (an estimated 243x speedup relative to one core).

"We're coordinating teams at UC Irvine, UC San Diego, the Santa Fe Institute, and Argonne National Laboratory so that after 130 years, we finally put comparative research on a solid footing of replicable scientific findings that are comparable from small to large and super-large datasets," said White.



K-cohesive levels of "southern women" neighborhood parties. K varies from 1 to 4 (connected to 4-connectivity). Image courtesy of James Moody and Douglas R. White

LOCAL IMPACT
STATEWIDE INFLUENCE
NATIONAL REACH

Our Local Impact

Top 10 REASONS why SDSC is of value to UC San Diego/San Diego

SDSC provides unique value to its home base of UC San Diego and its neighbors in the greater San Diego area. For the campus, SDSC is helping to assure UC San Diego's place among the great research universities in the world, on the cusp of the next great era of intellectual discovery: data-enabled science and engineering. For the local community, SDSC is working to establish research, education, and training partnerships with those needing the knowledge and skills to tackle problems posed by this new era, while providing expertise and resources to help protect the region's health and welfare.

As SDSC approaches its third decade, it does so as a leader in data cyberinfrastructure and technologies, and as a strong, collaborative institution with a clear and strategic focus to help solve the fundamental problems facing science and society. As such, SDSC provides value to UC San Diego and the greater San Diego region in the following 10 ways:

1. **Extensive research partnerships and collaborations**
   SDSC researchers collaborate with a large number of UC San Diego faculty and staff researchers on numerous projects led through SDSC and other campus units. Some 44 unique UC San Diego researchers in 23 UC San Diego departments/units submitted at least one extramural grant with SDSC's PIs over the past six years. Conversely, SDSC PIs collaborated on more than 50 extramural awards with a total exceeding $300 million to 21 other UC San Diego departments.

Geisel Library, UC San Diego

2. **A magnet for faculty recruitment**
   SDSC has helped UC San Diego attract some of the best and brightest faculty, who have partnered with SDSC to bring large-scale, competitive awards to the university. Recruitments have included Phil Bourne, who brought the Protein Data Bank to UC San Diego; J. Andrew McCammon, holder of the Joseph E. Meyer Chair of Theoretical Chemistry; Michael Norman, a distinguished professor of physics at UC San Diego, a globally recognized astrophysicist and SDSC director; and Shankar Subramaniam, developer of the Biology Workbench and partner in the UCSD/Nature Signaling Gateway, among others.

3. **Research cyberinfrastructure (RCI)**

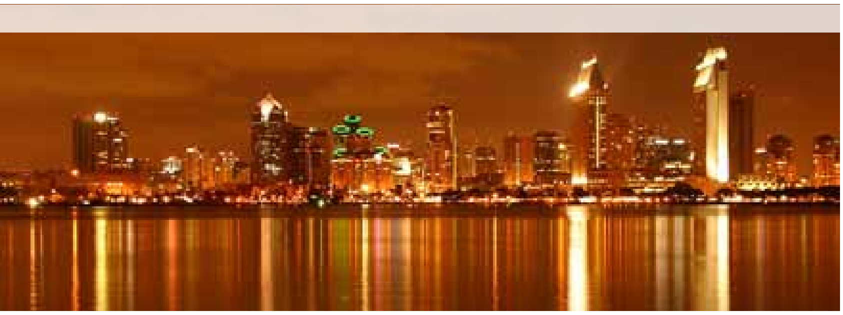
SDSC is a partner with several UC San Diego entities to offer research cyberinfrastructure to UC San Diego faculty and staff. The core elements consist of: co-location facilities housed in SDSC's data center; the SDSC Cloud, believed to be the largest academic cloud storage facility anywhere; digital curation and data services, provided by UCSD Libraries; the *Triton Shared Computing Cluster* (*TSCC*), serving researchers at UC San Diego and any of the other UC campuses as well as external academic, non-profit, and corporate users; and a research network, provided by the Administrative Computing and Telecommunications. (ACT).

4. **Pioneering design for high-performance computing**

SDSC houses *Gordon*, the first data-intensive supercomputer anywhere to use large amounts of flash-based memory—making it "the largest thumbdrive in the world." The result of a five-year $20 million grant from the NSF, *Gordon* has 300 trillion bytes of flash memory and 64 I/O nodes, making it ideal for data mining and exploration. SDSC also has a *Comet* sighting, thanks to a recent $12 million NSF award to launch in 2015 this new petascale supercomputer targeted to researchers in what's called the "long tail of science", which makes up about 99% of HPC users (see page 42).

5. **Technology innovation**

SDSC researchers have pioneered several significant software packages including ROCKS, which provides a blueprint for the construction of cluster computers for multiple laboratories around the globe; the Storage Research Broker, serving as "middleware" to hold together data cache sites for Big Data projects; iRODS, the open-source Integrated Rule-Oriented Data System, which represented a dramatic new approach to digital data management; and KEPLER, a scientific workflow automation system.



San Diego skyline. Image credit Tim McNew.

6. **The "Go To" place for "Big Data"**

SDSC is at the vanguard of the emerging 4th methodology of science known as data-enabled science and engineering, which seeks to apply massive data sets—commonly referred to as Big Data—to scientific discovery. Though the *Gordon* project puts SDSC at the vanguard, the Center's resources and expertise extend beyond the deployment and operation of this system for national users. SDSC's Big Data portfolio includes application team development, software development, user training and outreach, scientific collaborations, high-speed 100Gpbs (Gigabits per second) networking, training for today's academic and commercial researchers, and education for tomorrow's leaders.

7. **Network to the nation**

Since opening its doors nearly three decades ago, SDSC has conceived, nurtured, and raised multiple national partnerships and collaborations with individuals and communities across a wide spectrum of disciplines and fields of study. Today, SDSC is the only supercomputer center in the Western part of the nation participating in the National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE) as a service provider for advanced cyberinfrastructure services for the U.S. open research community. SDSC also provided the lynchpin for UC San Diego's partnership with the Open Science Grid (OSG), a multi-disciplinary consortium funded by the U.S. Department of Energy and the NSF.

8. **Education for the next generation**

SDSC is a founding partner in UC San Diego's Computational Science, Math and Engineering (CSME) program and serves as a "real world" training ground for more than 80 students each year, many of whom have taken advanced cyberinfrastructure skills to the private sector or to traditional academic environs. SDSC also invites students and teachers from the Greater San Diego region to experience computing at its highest levels during summer programs. In addition, the Center organizes a unique volunteer internship program for undergraduate students designed to provide valuable "on the job" experience in a cross-section of projects in computational and data research. Under a grant from The Parker Foundation, in partnership with select local middle schools and CONNECT's Entrepreneurs for Young Innovators program, SDSC is working to provide a new series of computer science workshops aimed at getting minority and female students involved in computing (see sidebar, page 35).

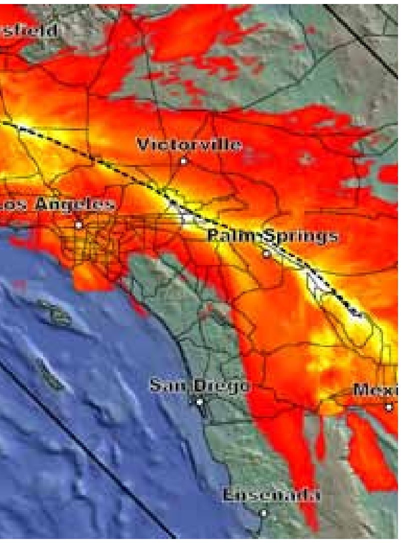


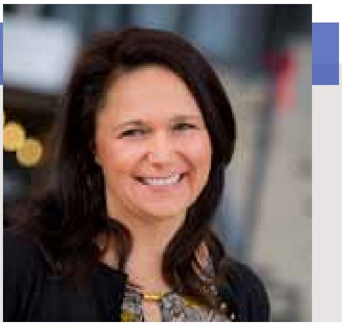
Image from a simulation of a magnitude 8 earthquake along the San Andreas Fault. Image credit: Amit Chourasia., SDSC

9. Outreach to the corporate community

SDSC has a long history of collaborating with and delivering value to industry, leveraging science and technology to deliver solutions for real-world problems. Among other programs, SDSC created the Predictive Analytics Center of Excellence (PACE) and its “boot camps” to accelerate research and education in predictive analytics for the academic and corporate communities. The Center for Large-scale Data Systems Research (CLDS) provides an opportunity for industry to collaborate with SDSC researchers on a large range of issues and challenges facing information-intensive organizations in this Big Data era. Finally, SDSC’s Industrial Partners Program (IPP) provides member companies with a framework for interacting with Center researchers and staff, exchanging information, receiving education and training, and developing collaborations.

10. Cyberinfrastructure to save lives and property from the “Big One”

Seismologists at SDSC, San Diego State University (SDSU), and the Southern California Earthquake Center (SCEC) at the University of Southern California (USC) recently created the largest-ever simulation of a Magnitude 8.0 M8) earthquake, along primarily the southern section of the San Andreas Fault. About 25 million people reside in that area, which extends as far south as Yuma, Arizona, and Ensenada, Mexico, and runs up through southern California as far north as Fresno. SDSC provided the high-performance computing and scientific visualization for the simulation. The research was selected as a finalist for the Gordon Bell prize, awarded annually for outstanding scientific achievement in high-performance computing applications. The work represents a major breakthrough in seismology both in terms of computational size and scalability. It also opens up new territory for earthquake science and engineering with the goal of reducing the potential for loss of life and property.



DEVELOPING A “TRULY INTELLIGENT” SMART GRID TO MANAGE ENERGY CONSUMPTION

Natasha Balac is director of SDSC’s Predictive Analytics Center of Excellence (PACE). Balac has developed a “sustainable communities” infrastructure proposal for downtown San Diego, in part to reduce power consumption.

Traditionally, electrical utility grids have been designed as a one-way street: to deliver power to the consumer. But with an ever-increasing demand to reduce energy costs and increase efficiencies, utility companies are now looking toward two-way street “smart grids” that not only deliver power, but also send back valuable information about how that power is being used to better manage these resources.

Toward that end, researchers with SDSC’s Predictive Analytics Center of Excellence (PACE) and UC San Diego are leveraging traditional meter-based power information, as well as weather, locale, time of day and other potentially significant factors, with predictive analytic tools to develop a “truly intelligent” smart grid for the UC San Diego’s 1200-acre campus. The results are expected to improve operating efficiency, lower costs, and reduce the campus’ overall carbon footprint generated by its 45MW peak load capacity.

“Achieving a truly smart energy infrastructure—for energy generation, distribution, and consumption—requires basic and advanced computing research and application,” said Natasha Balac, Director of SDSC’s PACE.

“The bidirectional flow of both electricity and information in the smart grid—consumption patterns in relation to the production, demand, environmental indicators and cost of energy—is extremely valuable when coupled with data analytics approaches,” she added.

A significant amount of campus building and energy data is collected from UC San Diego's smart grid. Data is currently collected from 84,000 distinct, independent data streams (with continuous measures at one minute intervals) from approximately 30 different campus buildings. Data streams, including real-time measurements and set points, are collected from various building management and control systems including HVAC systems, the central utility plant, electric power meters, photovoltaic panels, network model output data, weather stations, and even plug-in electric vehicles. This large, complex dynamic data set holds enormous potential for significant energy savings.

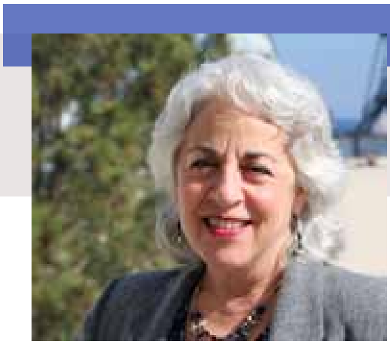
"UCSD's smart grid operators, as many others, have experienced that minimizing power consumptions does not necessarily reduce overall energy costs," said Balac. "To reduce the performance costs we are developing time series models that may be used to develop forward-thinking management policies. Integration of smart grid control, optimization and scheduling dramatically improves the controllers' ability to optimize indigenous resources, import energy, export surpluses and shed loads in a more optimal manner."

"The fundamental goal of this research is to significantly lower energy costs by applying predictive analytics algorithms on the existing master-controller-optimizer," she added.

A key part of the project is a data-mining technique called MineTool-TS (MineTool for Time Series data), enhanced by UC San Diego and SDSC researchers to capture time-lapse information for mining smart grid data.

Also participating in the project are Tamara Sipes and Homa Karimabadi, UCSD Jacobs School of Engineering; Nicole Wolter, Kenneth Nunes, and Robert Sinkovits, all from SDSC.

EXTENDING COMPUTER SCIENCE TO MINORITY AND FEMALE STUDENTS



Diane Baxter is associate director of education at SDSC, with a focus on introducing computational sciences and computational thinking skills to students and teachers in regional, national, and international settings.

SDSC, in a partnership with selected local middle schools and CONNECT's Entrepreneurs for Young Innovators program through a grant from the Parker Foundation, began a series of computer science workshops in 2013 aimed at getting more minority and female students involved in computing.

The workshops were designed to encourage middle school students to take a new Computer Science Principles course when they matriculate to high school. Participating students were from Granger Junior High School and National City Middle School in the Sweetwater Union High School District.

"San Diego's economy depends upon technology and innovation that impact nearly every job sector," said Diane Baxter, SDSC's associate director of education. "These workshops are all about getting more students, especially underrepresented ones, to learn the computational thinking skills that those jobs require. But most of all, we want them to engage in the fun and excitement of computing so they look forward to learning more as they continue their studies."

The middle school program, funded by a consortium of local start-up industries represented by CONNECT, dovetails with a larger program funded by the National Science Foundation (NSF) called ComPASS, for Computing Principles for All Students' Success. That program is focused on training teachers to instruct students in Computer Science Principles.

"Local industries clearly understand the need to engage the full diversity of the region's talent in their workforce," said Karen Winston, vice president of workforce development and STEM initiatives for CONNECT. "They want to be sure that all students are getting the preparation they need to be part of a technology-based, innovation-driven economy in the San Diego region."

Our Statewide Influence
Aligning with UC Principles

"Success in the Information Age can be measured by the precision, power, and breadth of the tools available to an organization, along with the knowledge and the creativity of the people who use them... Combined with the ability to integrate these components into world class 'cyberinfrastructure', SDSC has provided a foundation for research discoveries, education paradigms, and innovative business solutions. SDSC's impact on UC, the State of California, and beyond has been felt in many ways."

Governor Jerry Brown, December 8, 2011

Richard Moore is SDSC's deputy director, PI of the *Trestles* project, and co-PI of the new *Comet* supercomputer project.

Though SDSC was founded by the National Science Foundation (NSF) almost three decades ago in response to academic researchers nationwide seeking the latest supercomputing resources to solve "grand challenge" problems facing science and society, the Center's mission has evolved over time to include substantial work with, and for, the University of California's initiatives and researchers.

As part of that effort, SDSC has worked to align itself with three principles that "define the goals and purpose that drive and distinguish UC-wide research investments." Those principles are:

- Act as one system of multiple campuses to enhance UC's influence and advantage
- Promote efficient inter-campus collaborations and system-wide economies of scale
- Serve the State of California

Indeed, when SDSC became an organized research unit at UC San Diego in 1996, then Gov. Pete Wilson recognized the Center as a "cornerstone of California's vision for business and academic leadership in science and technology." Since then, SDSC not only has continued its highly successful national mission, the Center also has delivered value and prestige to UC and the State of California through numerous activities with a focus on its resources, services, and expertise in computational and data science. Some notable achievements include:

- **Research and discovery.** During the past five years, SDSC staff has collaborated on projects with more than 90 UC researchers across eight campuses. SDSC staff is currently participating in about 50 research grants and proposal collaborations across the UC system.

- **Funds through successful grant applications and research partnerships.** Over its almost 29-year history, SDSC revenues have exceeded $1B—a level of sustained funding matched by few academic research units in the country. SDSC has leveraged the operating funds it receives from UC Office of the President (UCOP) and UC San Diego and has returned six times the amount in sponsored research awards from FY2009 to FY2013.

- **High-performance and data-intensive computing.** Since its launch in 1985, more than 7,500 UC researchers have used SDSC's HPC systems. The Center currently provides compute and HPC storage resources to about 560 UC researchers, including 150 principle investigators across eight UC campuses.

- **Cost efficient colocation and "green" computing.** SDSC has made its 19,000 ft^2 data center available as a recharge-based colocation facility to more than 90 UC groups spanning eight campuses, with an estimated annual system-wide utility cost-savings that exceeds $500K.

- **Innovative cyberinfrastructure.** CI services, available at competitive rates to UC researchers, include the *Triton* computer resource which through 2012 provided more than a million core-hours to more than 600 users across eight UC campuses, and was used for teaching classes both at UC San Diego and UC Santa Barbara; in addition to the SDSC Cloud storage service, among the largest academic cloud storage system in the world.

- **Host to numerous critical databases.** SDSC houses numerous data sources of high impact including: the Protein Data Bank (protein structures); Medicaid data from the Centers for Medicare and Medicaid Services (CMS); the Library of Congress Chronopolis digital preservation system; Open Topography.org (LiDAR data); and the American Red Cross Safe and Well website, developed at SDSC in urgent response to the Hurricane Katrina disaster and the need to match missing people with family and friends. In 2010, SDSC worked with the office of California's CIO to launch the California Spatial Data Infrastructure Project, with 96 Terabyes (TB) of dedicated storage including the California Coastal Atlas which is being used to assess the impact of the sea level rise from climate change along the California coastline.

- **Bridge to empower the next generation.** Through its award-winning TeacherTech program, SDSC has trained more than 1,000 teachers in the San Diego region in science and technology, helping many underserved students to span the "digital divide" to the Information Age.

Aside from these current activities, SDSC has embarked on several initiatives for UC and the State of California, including:

- **Deployment of a massive computing system to support the needs of a larger and more expansive community of scientists and researchers, an activity sometimes referred to as the "long tail of science."** To help fill these needs and continue the Center's local and national focus on high-performance and data-intensive computing, SDSC is building the new *Comet* supercomputer, funded by NSF (see page 42).

- **Establishment of SDSC's brand in data science across UC through the creation of a new institute focused on data science and engineering.** To help fill the need to educate and train a new generation in data science, and continue the Center's innovations in research and discovery, SDSC is developing the Institute for Data Science and Engineering. This institute will develop and provide applied informatics hands-on experience and training across several UC campuses and UC San Diego. Some of the material will be suitable for an Introduction to Data Science undergraduate course offered to UC students system-wide, while other material could be used as part of a graduate curriculum, also across multiple graduate programs.

- **Building research collaborations focused on Big Data solutions for problems of high public interest for the benefit of UC and the State of California.** Under the Center's newly established "Data Initiatives" program, SDSC is reaching out for partnerships and collaborations on both the management and technical aspects of Big Data and other data-enabled applications. Projects under way included the BigData Top100 list, a community-based effort to establish the first global ranking for systems designed for Big Data applications, an effort that came as an offshoot of the Center for Large-scale Data Systems Research (CLDS), formed in 2012. SDSCs colo facility now stores all the cancer genomes from major projects funded by the National Cancer Institute, called CGHub, with ample room for expansion to house and provide secure access to all the genomic data from UC medical centers, and to provide the data analysis that will drive clinical genomic medicine in the future.

ANALYZING WILDFIRE BEHAVIOR

Three research organizations from UC San Diego, including SDSC, were awarded a multi-year National Science Foundation (NSF) grant in 2013 to build an end-to-end cyberinfrastructure to perform real-time-driven assessment, simulation, prediction, and visualization of wildfire behavior.

In addition to SDSC experts, the project—called WIFIRE—includes researchers from the California Institute for Telecommunications and Information Technology (Calit2) Qualcomm Institute, and the Mechanical and Aerospace Engineering (MAE) department with UCSD's Jacobs School of Engineering.

The WIFIRE CI (cyberinfrastructure) was designed to support an integrated system of wildfire analysis, with specific regard to changing urban dynamics and climate. The system integrates networked observations such as heterogeneous satellite data and real-time remote sensor data, with computational techniques in signal processing, visualization, modeling, and data simulations to provide a scalable method to monitor such phenomena as weather patterns that can help predict a wildfire's rate of spread.

The products of WIFIRE will be initially disseminated to project collaborators, including CAL FIRE, the U.S. Forest Service, and SDG&E covering academic, private, and government laboratories while providing value to emergency officials and first-responders, and in turn the general public.

WIFIRE will be available for use by government agencies in the future to save lives and property during wildfire events, test the effectiveness of response and evacuation scenarios before they occur, and assess the effectiveness of high-density sensor networks in improving fire and weather predictions.

Immersive technologies such as the NexCAVE at Calit2/Qualcomm Institute may play a central role in wildfire incident command center simulation and training. Photo by John Hanacek, Calit2/UC San Diego

"WIFIRE will be scalable to users with different skill levels using specialized web interfaces and user-specified alerts for environmental events broadcasted to receive before, during, and after a wildfire," said Ilkay Altintas, principal investigator for the WIFIRE project.

The WIFIRE CI encompasses the remote sensor network that is part of the High Performance Wireless Research and Education Network (HPWREN) project started at SDSC under NSF funding in 2000. HPWREN director and co-founder Hans-Werner Braun is a co-PI of WIFIRE; in addition to Larry Smarr, founding director of Calit2; and MAE Professor Raymond de Callafon.

"BIGDATA TOP100" TO BENCHMARK "BIG DATA"

Chaitan Baru, an SDSC Distinguished Scientist, was recently named the Center's Associate Director of Data Initiatives. Baru also directs SDSC's Center for Large-scale Data Systems Research (CLDS), focused on technology and technology management issues related to Big Data.

SDSC researchers are coordinating and providing the intellectual leadership toward the creation of a "BigData Top100" list, the first global ranking of its kind of Big Data systems, blending benchmarking approaches from high-performance computing, transaction processing, and database query processing. An initial board of directors has been formed to steer this activity, coordinated by Chaitan Baru, director of the Center for Large-scale Data Systems research (CLDS)—an industry-sponsored center of excellence created within SDSC to develop concepts, frameworks, analytical approaches, and systems solutions to address technical as well as technology management challenges facing information-intensive organizations in the era of Big Data. The board includes representatives from companies in California and beyond including Milind Bhandarkar, Pivotal; Dhruba Borthakur, Facebook; Eyal Gutkind, Mellanox; Jian Li, IBM; Raghunath Nambiar, Cisco; Meikel Poess, Oracle; and Tilman Rabl, University of Toronto.

SDSC ASSISTS RESEARCHERS IN NOVEL WILDLIFE TRACKING PROJECT

Amit Chourasia is a senior visualization scientist at SDSC and Visualization Services Group lead for the Center. Chourasia also is principle investigator for the SEEDME.org project.

A team including researchers from the U.S. Geological Survey (USGS) and the San Diego Zoo's Institute for Conservation Research has developed a novel methodology that for the first time combines 3D and advanced range estimator technologies to provide highly detailed data on the range and movements of terrestrial, aquatic, and avian wildlife species.

Relying on expertise from SDSC researchers, the team created highly detailed data sets and visualizations after they tracked three highly iconic but threatened species in the U.S., southwest China, and northeastern Australia: California condors, giant pandas, and dugongs—a large marine animal somewhat similar to the manatee.

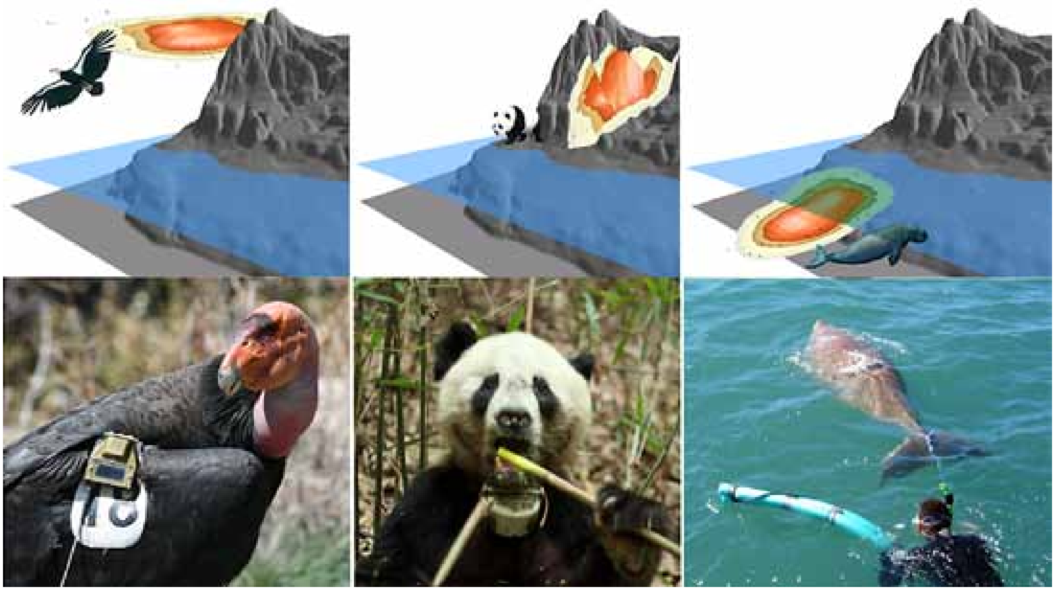
"We were able to speed up their software by several orders of magnitude," said Robert Sinkovits, director of SDSC's Scientific Applications Group, which helps researchers make optimal use of SDSC's larger supercomputers, including *Gordon* and *Trestles*. "In this case, calculations that had formerly taken four days to complete were finished in less than half an hour."

Gordon is being used to host the visualizations, and will make it easier for the software developers to explore the impact of algorithmic modifications on the quality of the solution. "Most importantly, we expect they will use our systems to solve other challenges that were previously considered to be intractable," said Sinkovits.

The visualization expertise was provided by Amit Chourasia, a senior visualization scientist at SDSC. "We made changes to write the data into a more compact format, which enabled swift output and ingestion," said Chourasia. "A key goal was to allow the experts to visualize the data directly on *Gordon* via remote access, as it is essential to minimize data movement and replication when data sizes grow. Currently, we're working to fuse data from various sources such as topography and climate to further aid the understanding of such habitats."

"Our collaborative research team harnessed the power of SDSC to fully exploit the increasing size and quality of 3D animal biotelemetry tracking and datasets," said James Sheppard, an ecologist with the San Diego Zoo's Institute for Conservation Research, and a member of the research team. "This gives us deeper insights into patterns of animal space-use, and informs strategies for the conservation management of endangered species and their habitats."

The study, called 'Movement-based Estimation and Visualization of Space Use in 3D for Wildlife Ecology and Conservation', was submitted to the *PLoS-ONE* online science journal. In addition to Sheppard, researchers for the study included Jeff Tracey, (USGS and lead author); Jun Zhu (University of Wisconsin, Madison); Fuwen Wei (Chinese Academy of Science, Beijing); Ronald Swaisgood (San Diego Institute for Conservation Research); and Robert Fisher (USGS, San Diego).



Example avian, terrestrial, and aquatic animal biotelemetry datasets and their spatial domains and home range contours. Left: California condor with a GPS biologger attached to its patagium. Center: A giant panda telemetered with a GPS collar. Right: A dugong fitted with a tail-mounted GPS biologger.

Our National Reach

Partnerships, Communities, and Collaborations

Historically, SDSC has conceived, nurtured, and raised multiple partnerships and collaborations with individuals, communities, and institutions across a wide spectrum of disciplines and fields of study across the nation.

These collaborations have yielded significant published papers and presentations at prestigious scientific meetings which, in turn, have offered new avenues for scientific discovery. To illustrate SDSC's research impact beyond UC San Diego, during the period 2007-2011, the Center received more than $80M in sub-awards from more than 50 non-UC San Diego research institutions, accounting for nearly 90 total awards. Similarly, SDSC allotted nearly $8M in sub-awards to 34 non-UC San Diego research partners during the same time period.

SDSC provides an organizational home for technology and science research and deployment for a wide range of projects to institutions in all regions of the country. Some of these activities include:

- **Partnership in the NSF's Extreme Science and Engineering Discovery Environment (XSEDE).** XSEDE, a partnership of 17 institutions, represents the most advanced, powerful, and robust collection of integrated advanced digital resources and services in the world. SDSC, the only supercomputer center participant on the West Coast, provides advanced user support and expertise for XSEDE researchers across a variety of HPC applications, in addition to support for the organization's central accounting database. SDSC also offers grid monitoring services through Inca, used by leading grid projects worldwide to identify, analyze, and troubleshoot user-level grid problems and failures.

- **Data-intensive HPC resources.** In 2011, *Trestles* came online to provide researchers from a diverse range of disciplines significant computing capabilities using flash-based memory, which can read data as much as 100 times faster than spinning disk, write data faster, and is more energy-efficient and reliable than standard disk technology. The following year, *Gordon* went into production as the first high-performance supercomputer to use large amounts of flash-based memory, making it the "largest thumbdrive in the world." The result of a five-year, $20 million grant from the NSF, *Gordon* has 300 trillion bytes of flash memory and 64 I/O nodes, making the system ideal for data mining and exploration, where researchers have to churn though tremendous amounts of data just to find a small amount of valuable information. In effect, *Gordon* is designed to do for scientific research what Google does for web searches.

- **OSG Members Gain Access to SDSC Compute Systems.** In June, UC San Diego and the Open Science Grid (OSG), a multi-disciplinary consortium funded by the U.S. Department of Energy and the National Science Foundation (NSF), announced a partnership under which campus researchers gained access to the OSG's fabric of Distributed High-Throughput Computing capabilities. The collaboration was designed to benefit researchers with high-throughput workloads commonly used in domains such as biomedical and life sciences, as well as the geosciences. The partnership means that UC San Diego is one of only a few research universities in the U.S. that is served by both OSG and NSF's XSEDE. It also means that members of the OSG would gain access to two key high-performance compute systems at SDSC, *Trestles* and *Gordon*.

- **Internet research for cybersecurity.** The Cooperative Association for Internet Data Analysis (CAIDA), based at SDSC, began work last year under a contract from the Department of Homeland Security's Science and Technology Directorate (DHS S&T) to help improve security in federal networks and across the Internet, while developing new and enhanced technologies for detecting, preventing, and responding to attacks on the nation's critical information infrastructure. Under this agreement, CAIDA will continue to grow its distributed Archipelago (or Ark) active-measurement infrastructure that currently consists of 62 monitors deployed in 29 countries on six continents. Researchers will use Ark to collect measurements from probes sent to all of the routed IPv4 prefixes on the Internet, and also will experiment with recently developed techniques that improve the efficiency and coverage of IP-level topology probing. The second phase of the project will focus on implementation techniques investigated during the first phase, while the third phase will include demonstration of these new technologies and systems in realistic operational environments, while continuing to add new and replace obsolete Ark monitors.

INNOVATIVE SCIENCE GATEWAYS

Nancy Wilkins-Diehr is an associate director of SDSC and co-director of XSEDE's Extended Collaborative Support Services. Her XSEDE responsibilities include providing user support for Science Gateways as well as education, outreach, and training.

Science gateways are used today to provide access to many of the tools used in cutting-edge research—telescopes, seismic shake tables, supercomputers, sky surveys, undersea sensors, and more. A single gateway can give thousands of users access to current, optimized versions of analysis codes at any time. SDSC has been a science gateways pioneer, and has provided innovative projects for XSEDE including the CIPRES Science Gateway (Cyberinfrastructure for Phylogenetic RESearch), representing almost 30 percent of all active XSEDE users (see page 25). During 2013, SDSC received a grant from the National Science Foundation (NSF) to create a software infrastructure for the Neuroscience Gateway. SDSC was also named the lead institute on a NSF planning grant for a Science Gateway Institute that would offer a complete range of services to develop domain-specific, user-friendly, Web-based portals and tools to build science gateways.

"COMET": FOR THE "LONG TAIL OF SCIENCE"

Last October, SDSC was awarded a $12 million grant from the NSF to deploy *Comet*, a new petascale supercomputer designed to transform advanced scientific computing by expanding access and capacity among traditional as well as non-traditional research domains. The new supercomputer, to be deployed in 2015 with another $12 million anticipated during the production phase, will be capable of an overall peak performance of nearly two petaflops (two quadrillion operations per second) and is designed to be part of an emerging cyberinfrastructure for what is called the "long tail of science", which targets a large number of modest-sized computationally based research projects. In effect, *Comet* will be the successor to SDSC's *Trestles* computer cluster, to be retired in 2014 after four years of service. As stated by SDSC Director Mike Norman: "*Comet* is all about high-performance computing for the 99 percent." *Comet* will be a Dell-based cluster based on next-generation Intel Xeon processors. Each node will be equipped with two of those processors, 128 GB (gigabytes) of traditional DRAM, and 320 GB of flash memory. Since *Comet* is designed to optimize capacity for modest-scale jobs, each rack of 72 nodes will have a full bisection InfiniBand FDR Interconnect, with a 4:1 bisection interconnect across the racks. *Comet* also will include some large-memory nodes, each with 1.5 TB of memory, as well as nodes with NVIDIA GPUs (graphics processing units). The GPU and large-memory nodes will target specific applications, such as visualization, molecular dynamics simulations, and de novo genome assembly.

HEALTHCARE INFORMATION TECHNOLOGIES

Dallas Thornton leads SDSC's Health IT division, focusing on delivering scale-out computing, data management, expertise, and support to partners at UC San Diego, other universities, and state and federal sponsors.

SDSC is building a growing Health IT center of excellence around several large grants in the expanding field of data-driven health care. Under an award from the Centers for Medicare and Medicaid Services (CMS) Medicaid Integrity Group (MIG) Data Engine, the Center is developing innovative solutions to detect fraud, waste, and abuse in the U.S. Medicaid programs. In addition, the group supports the National Children's Study San Diego County with secure cyberinfrastructure in its efforts to perform the largest longitudinal study of children and their development ever undertaken. Partnership in these awards extends SDSC's focus on Big Data in an increasingly socially important area. SDSC also is collaborating with David Haussler at UC Santa Cruz to store the Cancer Genomics Hub (CGHub), a large-scale data repository and user portal for the National Cancer Institute's cancer genome research programs, with the goal of targeting anti-cancer drugs to specific tumors and individual patients based on their genetic signatures.

PARTNERING FOR RESOURCES AND KNOW-HOW

PARTNERING for Resources and Know-how

SDSC INDUSTRY PARTNERS PROGRAM

Ron Hawkins is director of industry relations for SDSC and manages the Industry Partners Program, which provides member companies with a framework for interacting with SDSC researchers and staff to develop collaborations.

SDSC's focus on harnessing Big Data to advance scientific discovery has attracted numerous companies and external research institutes seeking to gain expertise or forge partnerships to manage vast amounts of data that could potentially create a competitive edge in research or commercialization.

Spanning areas such as biotech, civil engineering, health IT, transportation, and utilities, these organizations are educating themselves on everything from how to create sustainable data storage systems to learning about predictive analytics, or the process of using statistical techniques from modeling, data mining, and game theory to analyze current and historical facts to make predictions, as well as assess risks and opportunities, about future events.

In 2013, SDSC formally established an Industry Partners Program. The Center held its first annual Research Review for current and prospective industrial partners and affiliates as part of a broader strategy to foster such collaborations.

"Beginning with its roots in private industry (SDSC was founded by General Atomics Corporation in 1985), SDSC has a proud history of conducting 'applied R&D' and operating a production-quality computing infrastructure, making for natural synergies with industrial partners," said SDSC Director Michael Norman.

The Industry Partners Program is a way for SDSC to engage with smaller/startup companies, particularly across San Diego's innovation-driven economy. Organizations including the Council on Competitiveness and the National Center for Manufacturing Sciences have identified access to high-performance computing (HPC) as a key factor in maintaining America's manufacturing competitiveness among small- to medium-sized businesses. SDSC is also partnering with larger companies, such as San Diego Gas & Electric (SDG&E), which used the Center's HPC resources to fine-tune a high-resolution weather forecasting model to provide early warning for extreme fire dangers.

The advent of cloud computing has given companies another approach: renting compute capacity on an on-demand basis. While SDSC operates several medium- to large-scale compute clusters primarily configured for academic researchers, the Center now offers space-available access to companies, both local and distant, at market prices through its *Triton Shared Computing Cluster*, or *TSCC* (see page 9 for more details).

CLDS AND PACE: 'BIG DATA' CENTERS OF EXCELLENCE

The Industry Partners Program is closely linked to two other recent Big Data initiatives at SDSC: the Center for Large-scale Data Systems research (CLDS) and the Predictive Analytics Center of Excellence (PACE). Both organizations were recognized at a 2013 White House Office of Science and Technology Policy (OSTP) meeting for projects focused on accelerating collaborations in data-enabled science.

CLDS was established in 2012 to study the technology and management aspects of Big Data. It is facilitated by a National Science Foundation grant along with sponsorship and support from several companies confronting the Big Data phenomenon, including Seagate, Pivotal, NetApp, Brocade, Mellanox, and Cisco.

Chaitan Baru, an SDSC Distinguished Scientist and director of CLDS, was recognized at the White House OSTP event for launching a collaboration among industry, academia, and government to develop industry-standard, application-level benchmarks to evaluate hardware and software systems for Big Data applications. Called the BigData Top100 List, it is the first global ranking of its kind, and will include price/performance evaluations. The list will complement other widely used rankings of HPC systems such as the Top500 and Graph500.

"The tremendous growth in data has created the need for benchmarks to quantify system performance and price/performance for systems that perform big data tasks and applications," said Baru. "Experience shows that the existence of such benchmarks enables healthy competition among technology and solution providers, resulting in product innovation and evolving technologies."

Baru, who also serves as SDSC's associate director of data initiatives, sees CLDS playing a major role in creating sustainable industry partnerships that can be managed at the academic level. "My vision is for us to harness our capabilities to create a data science program that can help industry find a path through the data deluge. This requires the creation of a data infrastructure that can accompany a strong education and training program in data science."

PACE is a non-profit public educational organization dedicated to amplifying the power of predictive data analytics and developing a comprehensive, sustainable, and secure cyber-infrastructure. PACE's purpose is to foster collaboration and education among industry, government, and academia to find solutions to complex problems posed by the sheer amount of data being generated throughout science and society.

PACE leverages SDSC's data-intensive expertise in a new, multi-level curriculum designed to provide business and science enterprises the critical skills to design, build, verify, and test predictive data models. During 2013 PACE held several workshops called 'Data Mining Boot Camps', which attracted numerous industry participants. PACE Director Natasha Balac was recognized at the White House OSTP meeting for a project she is coordinating with Clean Tech San Diego and OSIsoft to develop a "sustainable communities" infrastructure for downtown San Diego, in part to reduce power consumption.

"We envision deploying a data infrastructure that connects physical systems such as those managing electricity, gas, water, waste, buildings, transportation and traffic," said Balac. "This project will enable the city of San Diego to use city-scale applications that will result in reduced electricity consumption and cost, while at the same time anticipating or uncovering grid instabilities, educating the public, and improving both the quality of life and economic development."

OSIsoft's software system will connect to and acquire significant volumes of detailed data streams which will be published in a cyber-secure, private cloud that is only accessible via signed and approved access mechanism protocols. SDG&E and UC San Diego have been beta-testing the OSIsoft software, and UC San Diego researchers are using the campus' microgrid system to analyze the data on the main SDG&E grid and the UC San Diego Smart Grid.

A key goal of the project is to develop a model for the collection and refinement of data that is applicable to other communities and applications. "Processes developed, as well as their results, will be published to help enable other communities on their own path to sustainability," added Balac.

CASE STUDY: IRRIGATION MAKER TURNS TO SDSC TO HELP KEEP ITS PRODUCTS FLOWING

As California practices greater environmental stewardship, the use of recycled and unfiltered water for irrigation purposes has grown significantly. With that has come a new challenge for companies such as Hunter Industries, one of the world's leading makers of water-efficient irrigation products for residential, commercial, and golf course applications. Fine silt and other debris often exist in unfiltered water sources such as lakes, rivers, and wells, causing internal components of sprinklers to clog and malfunction over time.

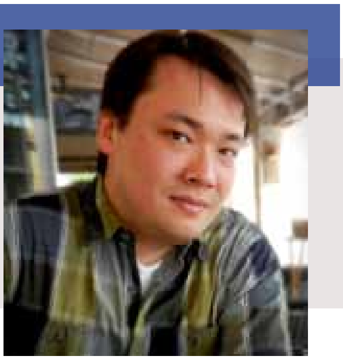
When Hunter's product designers wanted to enhance the performance of their existing line of sprinkler heads, they turned to SDSC to help them develop increasingly fine-resolution CFD (computational fluid dynamics) simulations. Using SDSC's *Gordon* supercomputer, the project provisioned a one-terabyte Windows virtual machine capable of tackling large modeling and simulation problems that do not scale well on conventional clusters.

"One of the largest challenges we faced was our limited knowledge of HPC systems," said Hunter Industries Vice President Gene Smith. "As industry engineers with a focus on manufacturing and smaller-scale simulations, we had little or no experience with Linux or the complexities of HPC cluster configurations. We found that while the CFD-solver itself scaled well across the computing cluster, every step up in resolution took significantly more time for mesh generation, dramatically slowing the process."

Courtesy of Hunter Industries

Although various technical issues and the timeframe of the collaboration precluded running actual simulations, the project was an anecdotal validation of the case made by the Council on Competitiveness regarding the role of HPC in enhancing the competitiveness of small and medium-sized manufacturing enterprises.

"We can now better determine the precise level of mesh refinement that balances set-up time with computing costs, while delivering timely, precise results," said Smith. "HPC will certainly be a valuable tool for us going forward as we increase our reliance on CFD simulation to reduce costs and time in prototyping and design."



SPEEDING BIG DATA GENOMIC ANALYSIS OF RHEUMATOID ARTHRITIS

Glenn Lockwood is a user service consultant for SDSC, providing support to users of the Center's high-performance computing resources to make supercomputing less obtuse and more accessible to researchers and the public.

Janssen Research and Development, LLC (Janssen), in collaboration with SDSC and the Scripps Translational Science Institute (STSI), recently launched a project to conduct whole-genome sequencing of 438 patients with rheumatoid arthritis to better understand the disease, as well as explore genetic factors of patient response to a biologic therapy discovered, developed, and currently marketed by Janssen in the United States.

The large-scale sequencing study—performed with the aid of SDSC compute and data resources, *Gordon* and *Data Oasis* —was designed not only to identify specific genetic variants in these patients that would make them more or less likely to respond to treatment, but also to help researchers wanting to search the entire genome for other correlations to disease.

The initial genetic analysis for the project was done by Kristopher Standish, a UC San Diego graduate student, working under Nicholas Schork at the Scripps Translational Science Institute. The Big Data computational expertise came from Glenn K. Lockwood, Wayne Pfeiffer, and others at SDSC.

The analysis started with 50 terabytes (TB) of human genome data from 438 rheumatoid arthritis patients consisting of "compressed reads"—strings of 100 DNA bases (A, adenine; G, guanine; C, cytosine; and T, thymine) generated as outputs by sequencers. The subsequent analysis included a complex 14-step pipeline using community codes BWA, SAMtools, Picard, and GATK to map against a reference genome and then identify statistically significant genetic variants.

The computational parallelism as well as the memory and input/output (I/O) requirements varied throughout the pipeline, which necessitated careful orchestration of the steps to achieve fast and efficient processing. *Gordon*, the nation's first data-intensive supercomputer with large quantities of flash memory, provided an innovative compute environment for this effort.

"One problematic step involved a massive sort of records in intermediate files," said Pfeiffer. "This was accommodated by using two 'BigFlash' nodes of *Gordon*, each with 4.4 TB of usable flash memory. This allowed 3.5 TB of data to be sorted in each node using all 16 of its cores running in parallel."

At its peak, the project used about 5,000 cores, or roughly 30 percent of *Gordon*, and nearly 350 TB of *Data Oasis*. Said Lockwood: "The bulk of the analysis was completed in six weeks (including learning time on *Gordon*) using more than 300,000 core hours of computer time, compared to more than four years of 24/7 compute time on an 8-core workstation.“

"We were very pleased with these results and believe this project demonstrates the possibilities for future collaborations in need of fast and efficient Big Data genetic analysis," Pfeiffer added.

DYS DYS DYS DYS DYS DYS DYS DYS DYS
389II 390 391 392 393 426 438 439 441
15 19 14 16 13 15 19 1

HPWREN TO THE RESCUE: RESEARCH NETWORK EVOLVES INTO PUBLIC SAFETY ASSET

Hans-Werner Braun is an SDSC research scientist who helps manage the Area Situational Awareness for Public Safety Network (ASAPnet), which provides broadband and Internet connectivity to about 60 fire stations in remote parts of San Diego County. Braun, along with UC San Diego Scripps Institution of Oceanography Seismologist Frank Vernon, co-founded HPWREN in 2000.

When a fast-moving blaze burned more than 7,000 acres near Mount Laguna in August 2013, firefighters were able to monitor its spread and respond accordingly by relying on a high-speed data transmission network made possible in part by UC San Diego's High-Performance Wireless and Research Education Network (HPWREN).

By the time that wildfire was contained, more than 10,000 people, in addition to rescue crews, had accessed HPWREN's camera images, once again demonstrating the network's value as a public safety asset throughout greater San Diego. HPWREN cameras aided firefighters in 2003 and 2007, when devastating wildfires swept through large parts of San Diego county.

Today, with the assistance of San Diego Gas & Electric (SDG&E) and other partners, HPWREN, via a program called the Area Situational Awareness for Public Safety Network or ASAPnet, provides broadband and Internet connectivity to about 60 fire stations in remote parts of San Diego county. The backbone of ASAPnet is directed by Hans-Werner Braun, a research scientist at SDSC who, along with Frank Vernon, a seismologist with the UC San Diego Scripps Institution of Oceanography, founded HPWREN in 2000.

Throughout its existence, HPWREN has attracted a variety of users that rely on its high bandwidth connectivity to remote areas for purposes such as studying earthquakes, monitoring wildlife, conducting educational activities, and monitoring firefighting operations. In 2011, HPWREN transitioned from

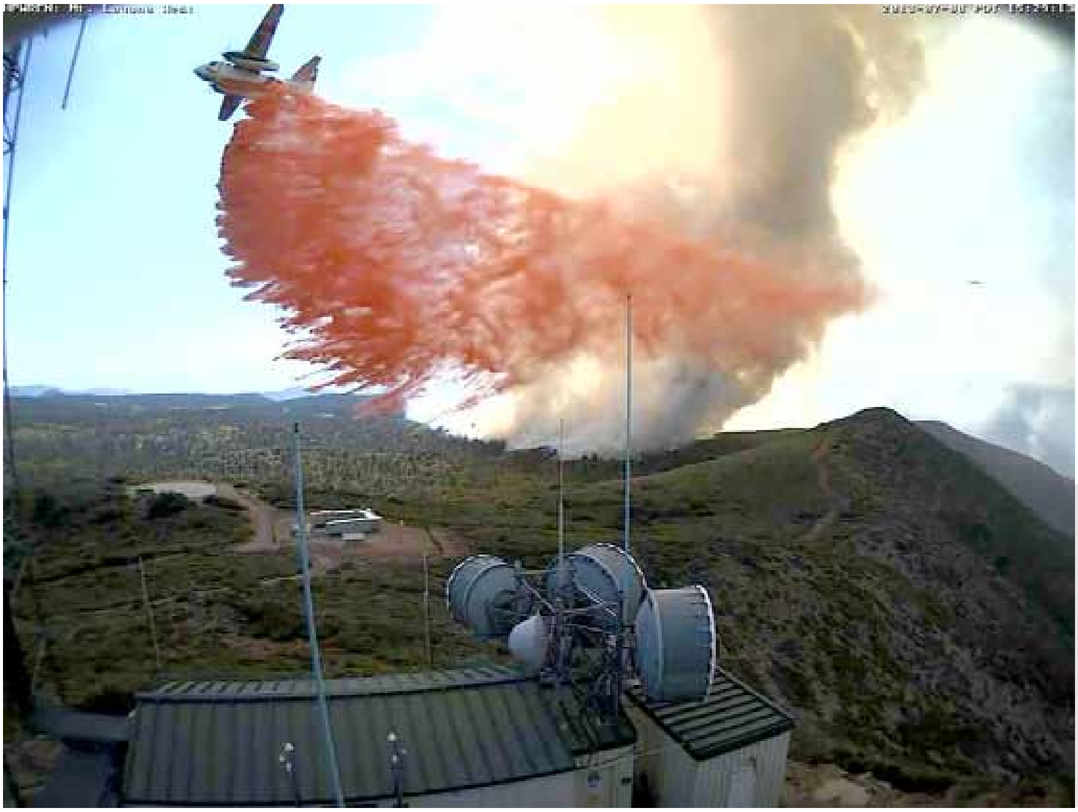
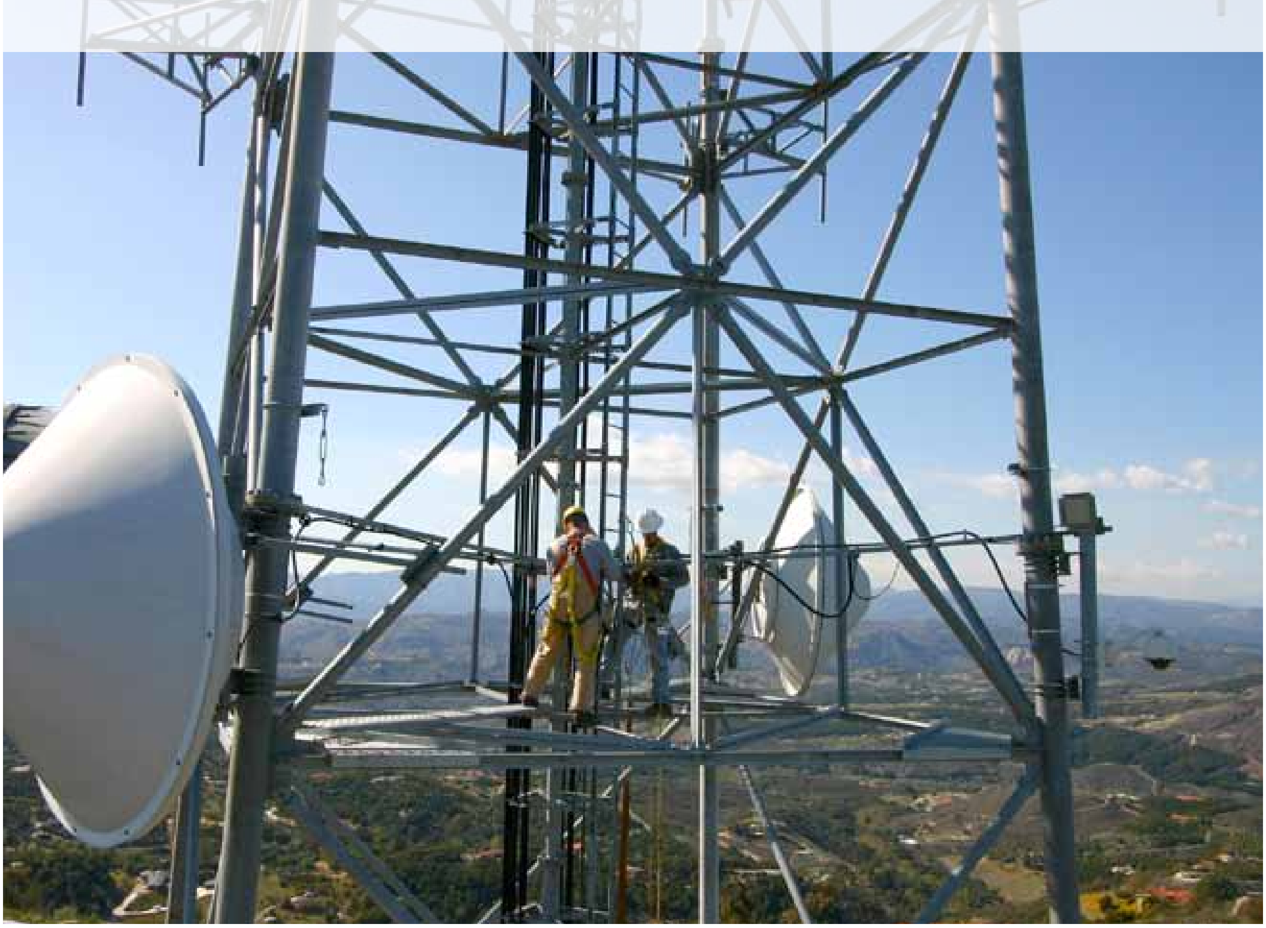


Image of the July 2013 Chariot Fire on Mt. Laguna near San Diego, via a stationary HPWREN camera provided by SDG&E. Courtesy of HPWREN/SDSC.



Pablo Bryant and Mark VanScoy, from San Diego State U, partnering with HPWREN on installation work at an HPWREN backbone site. Image credit: Hans-Werner Braun.

an NSF-supported funding model to a collaborative partnership funded by its user community. In recent years, through ASAPnet and other initiatives, HPWREN has taken on an increasing role in public safety.

"It took us 12 years to get to 15 fire stations linked to the network, but just one additional year to get to 60," said Braun, who was part of a joint team to conduct tasks such as aligning mountaintop wireless antennas with those on remote fire stations.

Aware of the work being done through previous public-safety collaborations with HPWREN, San Diego County Supervisor Ron Roberts in 2012 brought officials with CAL FIRE and the San Diego County Fire Authority together with SDG&E to formally establish ASAPnet and further cement HPWREN's role as a public safety asset for San Diego County.

Braun judges the ASAPnet partnership a success. "This collaboration has also allowed for the provisioning of many more environment-observing cameras with capabilities beyond the previously used cameras (such as near-infrared sensing), as well as several more weather stations that can provide firefighters with up-to-the-second real-time wind data," according to Braun. "Taken together, this has substantially enhanced the capabilities of HPWREN and our ability to effectively and immediately have a positive impact on the public at large, especially in such rapidly changing and dangerous situations."

"SDG&E is proud to be part of a collaborative effort that benefits the entire community—specifically, fire agencies and first responders," said Michael R. Niggli, SDG&E's president and chief operating officer. "The investments we are making in technology such as HPWREN are geared primarily to improving SDG&E's overall situational awareness, and it is deeply satisfying to know the benefits are multiplied across our entire service area."

"Our ASAPnet collaboration is just one example of SDSC's efforts to partner with industry and government leaders in a variety of areas to create programs that benefit society on many levels," added SDSC Director Michael Norman. "These partnerships can pay long-term dividends to both local communities, as well as communities around the world that can use the latest infrastructures and advanced technologies to help solve everyday challenges."

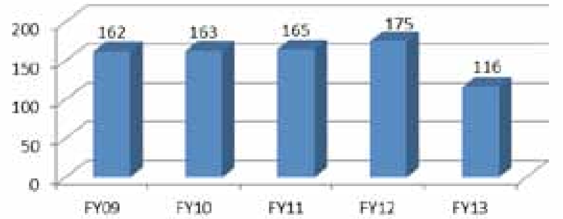
FACTS & FIGURES

Proposal Success Rate

	FY09	FY10	FY11	FY12	FY13
Proposals Submitted	81	119	92	116	94
Proposals Funded	43	44	44	51	40
Success Rate	53%	37%	48%	44%	43%

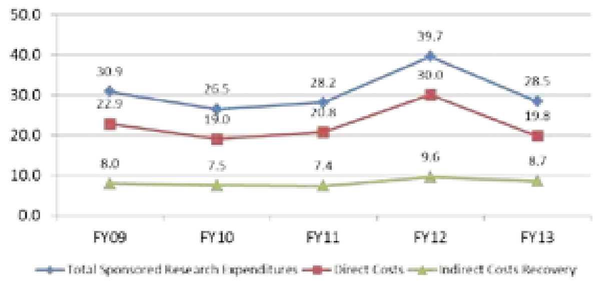
In perhaps the most competitive landscape for federal funding in the last two decades, SDSC's overall success rate on federal proposals is currently 45%, compared to a national average of roughly 15% for computer science and engineering proposals at the National Science Foundation.

Number of Sponsored Research Awards



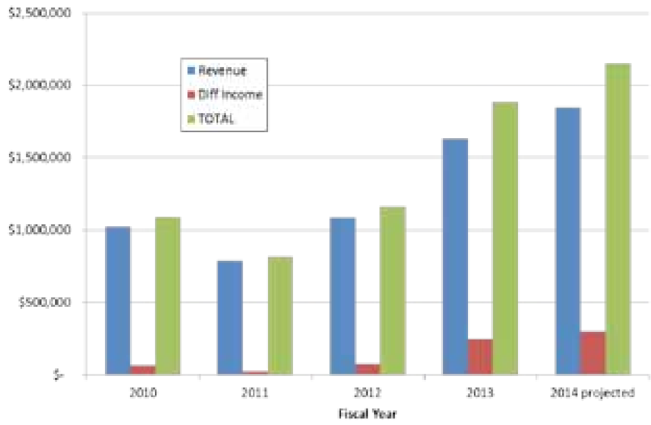
FY13 decline represents a shift to fewer awards with larger budgets.

Sponsored Project Expenditures ($M)

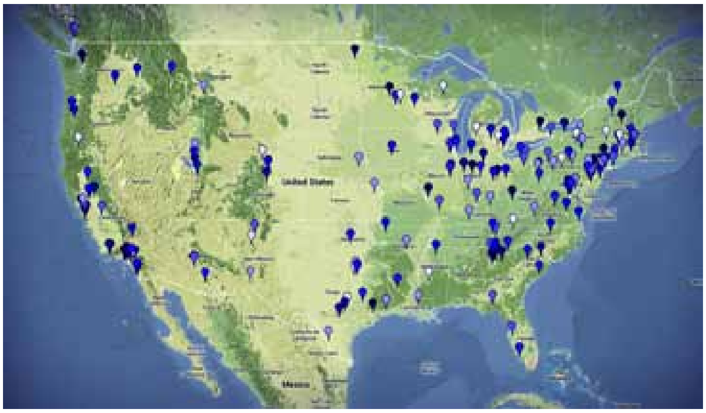


Apart from the extraordinary research impact of SDSC collaborations and partnerships, a quick look at the fiscal impact of these collaborations is impressive. During its 27-year history, SDSC revenues have exceeded $1 billion, a level of sustained funding matched by few academic research units in the country. In the above graph, FY12 expenses are higher due to a $10M hardware purchase for NSF's *Gordon* Supercomputer. While the graph shows a decline in FY13 expenditures due to the FY12 acquisition, the SDSC award base has remained stable.

Industry Revenue Data 2010-2014 (projected)



Geographical Distribution of National Users of SDSC HPC Resources

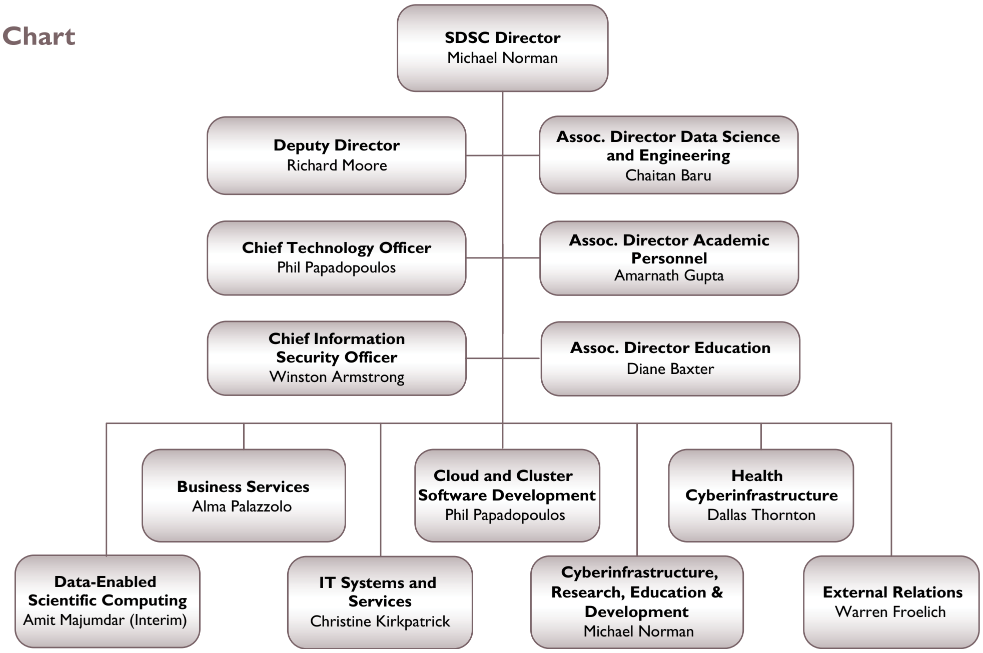


A total of 957 unique users from around the world accessed SDSC's HPC resources (*Gordon* and *Trestles*) during FY2013. Of these users, 916 were based in the United States. The adjacent map displays a geographic disbursements of users from different cities across the U.S.

On *Gordon*, a total of 85,423,681 service units (SUs) were used, 94% of which were charged against XSEDE accounts. On *Trestles*, a total of 58,740,999 SUs were used, 98% of which were charged against XSEDE accounts.

ORGANIZATION & LEADERSHIP

SDSC Org Chart



SDSC Director
Michael Norman

Deputy Director
Richard Moore

Assoc. Director Data Science and Engineering
Chaitan Baru

Chief Technology Officer
Phil Papadopoulos

Assoc. Director Academic Personnel
Amarnath Gupta

Chief Information Security Officer
Winston Armstrong

Assoc. Director Education
Diane Baxter

Business Services
Alma Palazzolo

Cloud and Cluster Software Development
Phil Papadopoulos

Health Cyberinfrastructure
Dallas Thornton

Data-Enabled Scientific Computing
Amit Majumdar (Interim)

IT Systems and Services
Christine Kirkpatrick

Cyberinfrastructure, Research, Education & Development
Michael Norman

External Relations
Warren Froelich

Executive Team

Chaitanya Baru
Assoc. Director Data Science and Engineering

Warren Froelich
Division Director External Relations

Ronald Bruce Hawkins
Director Industry Relations

Christine Kirkpatrick
Division Director IT Systems and Services

Richard Moore
Deputy Director

Michael Norman
SDSC Director

Alma Palazzolo
Division Director Business Services

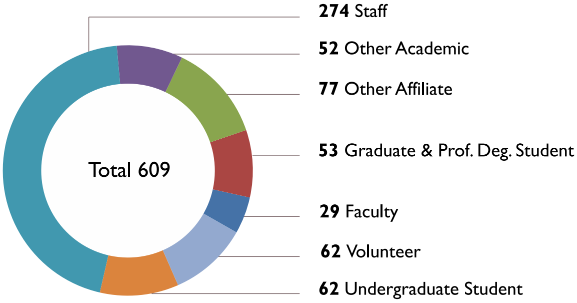
Philip M. Papadopoulos
Division Director Cloud & Cluster Software Development

Executive Committee

UC SAN DIEGO
Sandra Brown
Mark Ellisman
Michael Holst
J. Andrew McCammon
John Orcutt
Tajana Rosing
Nicholas Schork
Brian Schottlaender
Robert Sullivan
Susan Taylor

SDSC
Chaitanya Baru
Philip Bourne
Warren Froelich
Ronald Bruce Hawkins
Christine Kirkpatrick
Richard Moore
Michael Norman
Alma Palazzolo
Philip M Papadopoulos
Dallas Thornton

SDSC Census FY2013



RESEARCH EXPERTS

SDSC Computational Scientists

Laura Carrington, Ph.D.
Director, Performance, Modeling, and Characterization Lab, SDSC
Principal Investigator, Institute for Sustained Performance, Energy, and Resilience (DoE)
HPC benchmarking, workload analysis
Application performance modeling
Energy-efficient computing
Chemical engineering

Dong Ju Choi, Ph.D.
Senior Computational Scientist, SDSC
HPC software, programming, optimization
Visualization
Database and web programming
Finite element analysis

Yifeng Cui, Ph.D.
Director, High-performance GeoComputing Laboratory, SDSC
Principal Investigator, Southern California Earthquake Center
Senior Computational Scientist, SDSC
Adjunct Professor, San Diego State University
Earthquake simulations
Parallelization, optimization, and performance evaluation for HPC
Multimedia design and visualization

Andreas Goetz, Ph.D.
Co-Director, CUDA Teaching Center
Quantum Chemistry
Molecular Dynamics
ADF and AMBER developer
GPU accelerated computing

Glenn K. Lockwood, Ph.D.
SDSC User Services Consultant
High-performance computing
Materials science
Molecular dynamics
Large-scale genomic analysis
XSEDE

Amit Majumdar, Ph.D.
Interim Division Director, Data Enabled Scientific Computing, SDSC
Associate Professor, Department of Radiation Medicine and Applied Sciences, UCSD
Algorithm development
Code optimization
Code profiling/tuning
Science Gateways
Nuclear engineering

Michael Norman, Ph.D.
Director, San Diego Supercomputer Center
Distinguished Professor, Physics, UCSD
Director, Laboratory for Computational Astrophysics, UCSD
Computational astrophysics

Dmitri Pekurovsky, Ph.D.
Member, Scientific Computing Applications group, SDSC
Optimization of software for scientific applications
Performance evaluation of software for scientific applications
Parallel 3-D Fast Fourier Transforms
Elementary particle physics (lattice gauge theory)

Wayne Pfeiffer, Ph.D.
Distinguished Scientist, SDSC
Supercomputer performance analysis
Novel computer architectures
Bioinformatics

Bob Sinkovits, Ph.D.
Scientific Applications Lead
Gordon applications
Data-intensive high-performance computing
Computational physics and fluid dynamics
Bioinformatics
Relationship databases
Compute clusters systems administration

Mahidhar Tatineni, Ph.D.
User Support Group Lead, SDSC
Research Programmer Analyst
Optimization and parallelization for HPC systems
Aerospace engineering

Igor Tsigelny, Ph.D.
Research Scientist, SDSC
Research Scientist, Department of Neurosciences, UCSD
Computational drug design
Personalized cancer medicine
Gene networks analysis
Molecular modeling/molecular dynamics
Neuroscience

Rick Wagner, Ph.D. candidate
High-performance Computing Systems Manager
Large-scale Linux-based high-performance computing clusters
Cyberinfrastructure systems architecture and design
Computational astrophysics

Ross Walker, Ph.D.
Director, Walker Molecular Dynamics Lab
Co-Director, CUDA Teaching Center
Director, Intel Parallel Computing Center
Molecular dynamics
Quantum chemistry
GPU accelerated computing

Nancy Wilkins-Diehr, M.S.
Co-Principal Director, XSEDE at SDSC
Co-Director for Extended Collaborative Support, XSEDE
Science gateways
User services
Aerospace engineering

SDSC Data Scientists

Ilkay Altintas, Ph.D
Director, Workflows for Data Science (WorDS) Center of Excellence
Director, Scientific Workflow Automation Technologies (SWAT) Laboratory
Lecturer, Computer Science and Engineering @ UCSD
Assistant Research Scientist, SDSC
Scientific workflows
Big Data applications
Distributed computing
Reproducible science
Kepler Scientific Workflow System

Michael Baitaluk, Ph.D.
Assistant Research Scientist, SDSC
Principal Investigator, Biological Networks, SDSC
Scientific data modeling and information integration
Gene networks
Systems and molecular biology
Bioinformatics

Natasha Balac, Ph.D.
Director, Predictive Analytics Center of Excellence, SDSC
Director of Data Application and Services, SDSC
Data mining and analysis
Machine learning
Predictive analytics
Data-intensive computing
Big Data analytics

Chaitan Baru, Ph.D.
SDSC Distinguished Scientist
Director, Center for Large-scale Data Systems research (CLDS), SDSC
Associate Director, Data Science and Engineering, SDSC
Data management
Large-scale data systems
Data analytics
Parallel database systems

Amit Chourasia, M.S.
Senior Visualization Scientist, SDSC
Lead, Visualization Services Group
Principal Investigator, SEEDME.org
Visualization and computer graphics
Ubiquitous Sharing Infrastructure

Alberto Dainotti, Ph.D.
Research Scientist, CAIDA (Cooperative Association for Internet Data Analysis)
Internet measurements
Traffic analysis
Network security
Large-scale internet events

Amogh Dhamdhere, Ph.D.
Assistant Research Scientist, CAIDA (Cooperative Association for Internet Data Analysis)
Internet topology and traffic
Internet economics
IPv6 topology and performance
Network monitoring and troubleshooting

kc claffy, Ph.D.
Director/Principal Investigator, CAIDA (Cooperative Association for Internet Data Analysis), SDSC
Adjunct Professor, Computer Science and Engineering, UCSD
Internet data collection, analysis, visualization
Internet infrastructure development of tools and analysis
Methodologies for scalable global Internet

Amarnath Gupta, Ph.D.
Associate Director, Academic Personnel
Director of the Advanced Query Processing Lab, SDSC
Co-principal investigator, Neuroscience Information Framework (NIF) project, Calit2
Bioinformatics
Scientific data modeling
Information integration and multimedia databases
Spatiotemporal data management

Mark Miller, Ph.D.
Principal Investigator, Biology, SDSC
Principal Investigator, CIPRES gateway, SDSC/XSEDE
Principal investigator, Research, Education and Development group, SDSC
Structural biology/crystallography
Bioinformatics
Next-generation tools for biology

Dave Nadeau, Ph.D.
Senior Visualization Researcher, SDSC
Data mining
Visualization techniques
User interface design
High-dimensionality data sets
Software development
Audio synthesis

Philip M. Papadopoulos, Ph.D.
Chief Technology Officer, SDSC
Division Director, Cloud and Cluster Software Development, SDSC
Associate Research Professor (Adjunct), Computer Science, UCSD
Rocks HPC cluster tool kit
Virtual and cloud computing
Data-intensive, high-speed networking
Optical networks/OptIPuter
Prism@UCSD

Julia V. Ponomarenko, PhD.
Senior Research Scientist, SDSC
Structural bioinformatics
Bioinformatics & systems biology
Immunoinformatics & Immune Epitope Database (IEDB)
Biological data integration & integrative analysis

Andreas Prlić, Ph.D.
Senior Scientist / Associate Project Scientist, Protein Data Bank
Bioinformatics
Structural biology
Computational biology
Protein Data Bank

Peter Rose, Ph.D.
Scientific Lead, Protein Data Bank
Structure-based drug design
Bioinformatics
Computational biology
Protein Data Bank

Karen Stocks, Ph.D.
Specialist, SIO/SDSC
Ocean and biodiversity informatics
Metadata

Jianwu Wang, Ph.D.
Assistant Project Scientist, SDSC
Lecturer, Computer Science & Engineering, UCSD
Scientific workflow automation
Data-intensive computing

Hans-Werner Braun
Research Scientist, SDSC
Adjunct Professor, College of Sciences, SDSU
Director/PI, High Performance Wireless Research and Education Network (HPWREN)
Internet infrastructure, measurement/analysis tools
Wireless and sensor networks
Internet pioneer (PI, NSFNET backbone project)
Multi-disciplinary and multi-intitutional collaborations

Ilya Zaslavsky, Ph.D.
Director, Spatial Information Systems Laboratory, SDSC
Spatial and temporal data integration/analysis
Geographic information systems
Hydrology
Spatial management infrastructure

Andrea Zonca, Ph.D.
HPC Applications Specialist
Data-intensive computing
Data visualization
Cosmic microwave background
Python development

SDSC

San Diego Supercomputer Center
University of California, San Diego
9500 Gilman Drive MC 0505
La Jolla, CA 92093-0505

www.sdsc.edu
twitter/SDSC_UCSD
facebook/SanDiegoSupercomputerCenter