

SAN DIEGO SUPERCOMPUTER CENTER at UC SAN DIEGO



# GATEWAY to DISCOVERY

Tackling Today's Grand Research Challenges



**SDSC**

ANNUAL REPORT FY2016/17

## SAN DIEGO SUPERCOMPUTER CENTER

As an Organized Research Unit of UC San Diego, SDSC is considered a leader in data-intensive computing and cyberinfrastructure, providing resources, services, and expertise to the national research community, including industry and academia. Cyberinfrastructure refers to an accessible, integrated network of computer-based resources and expertise, focused on accelerating scientific inquiry and discovery. SDSC supports hundreds of multidisciplinary programs spanning a wide variety of domains, from earth sciences and biology to astrophysics, bioinformatics, and health IT. SDSC's petascale *Comet* supercomputer continues to be a key resource within the National Science Foundation's XSEDE (Extreme Science and Engineering Discovery Environment) program.

## 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

[info@sdsc.edu](mailto:info@sdsc.edu)  
[www.sdsc.edu](http://www.sdsc.edu)

Jan Zverina  
Division Director, External Relations  
[jzverina@sdsc.edu](mailto:jzverina@sdsc.edu)  
858-534-5111

## GATEWAY TO DISCOVERY

### SDSC Annual Report FY2016/17

(PDF version available online at the SDSC website)

EDITOR:  
Jan Zverina

CO-EDITOR:  
Warren Froelich

CONTRIBUTORS:  
Warren Froelich, Julie Gallardo, Ron Hawkins,  
Nieves Rankin, Wayne Pfeiffer, Susan Rathbun,  
Bob Sinkovits, Shawn Strande, Ben Tolo, Nancy Wilkins-  
Diehr, Nicole Wolter, Jan Zverina

CREATIVE DIRECTOR:  
Ben Tolo

DESIGN:  
Ben Tolo

All financial information is for the fiscal year ended June 30, 2017. Any opinions, 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.

© 2017 The Regents of the University of California



Director's Letter  
Charting the Course  
Page 2



SDSC's "Pi Person" of the Year  
Page 4

## SDSC's Computational Resources

Page 6-14

## Impact and Influence

Page 25

## Science Highlights

Page 15



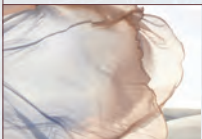
16-17

Biodiversity



18-19

Artificial Intelligence  
Quality of Life



20-21

Materials Engineering



22-24

Human Health  
Life Sciences



26-31

Local Impact & Influence



32-35

State Impact & Influence



36-42

National Impact & Influence

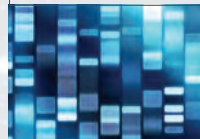
## Focused Solutions/Applications

Page 43



46

Advanced, Versatile Computing Systems



47

Life Science Computing & Applications



48-53

Data Science Platforms & Applications

## Industry

Page 54-57

## Facts & Figures

Page 58-61



## DIRECTOR'S LETTER

# CHARTING THE COURSE

Today, data-enabled science and life science research are based on computationally-based research, and would not be able to flourish without this effort.

SDSC devoted a good part of 2016 and 2017 to develop a comprehensive strategic plan that charts our course for the next three-to-five years. With this plan we have embarked on a path that builds on SDSC's core strengths as a national resource for advanced computing and data-enabled science that aligns with UC San Diego's vision and mission, described by Chancellor Khosla as being an "experimental campus" that follows an "unconventional tradition of non-tradition." Our priorities also support the goals of the University of California system and meet the myriad scientific challenges of the national, state, and local research communities.

To recap, SDSC's three key strategic priorities are:

- **Versatile Computing**  
This priority builds on SDSC's core strengths in advanced computing, which are now essential to nearly every field of research, with demand for computing far outstripping availability of resources. It includes our leadership in establishing sustainable science gateways, or application-specific, web-based interfaces for accessing supercomputers that let users focus on their science

instead of the systems. It also includes high-throughput computing as a critical tool in the analysis of data sets from large-scale experiments and cloud computing, which has become an important element of today's research landscape.

- **Data Science and Engineering**  
Data-driven science has emerged as one of the single most transformative forces for discovery. Large scientific experiments, gene sequencers, machine and human sensors, and the Internet of Things all present those responsible for building cyberinfrastructure tools with major technological, applications, and workforce challenges in developing end-to-end solutions for a wide range of critical research.
- **Life Sciences Computing**  
This priority aligns with national efforts such as the National Science Foundation's (NSF) 'Understanding the Brain' initiative. It also supports more efficient inter-campus collaborations such as UC San Diego's strategic priority of "exploring the basis of human knowledge and creativity", while providing UC-wide economies-of-scale. SDSC is well-positioned to foster new collaborations with the life sciences research community to develop advanced platforms and services.

### SDSC's VISION STATEMENT

TO DELIVER LASTING IMPACT ACROSS THE GREATER SCIENTIFIC COMMUNITY BY CREATING END-TO-END COMPUTATIONAL AND DATA SOLUTIONS TO MEET THE BIGGEST RESEARCH CHALLENGES OF OUR TIME.

Let me emphasize that these three priorities not only are closely aligned, but are firmly intertwined with each other. Today, data-enabled science and the life sciences are based on computationally-based research, and would not be able to flourish without this effort. Moreover, SDSC's standing as a well-regarded national supercomputing center, along with the Center's commitment to fulfilling its national advanced cyberinfrastructure mission, provides significant benefits to UC San Diego, the UC system, and our industry partners. One such example is the recent NSF Campus Cyberinfrastructure (CC\*) award to enhance bioinformatics computing capabilities for UC San Diego researchers using the campus' *Triton Shared Computing Cluster* housed at SDSC, which you can read more about on page 11.

In short, many such awards and collaborations would not exist without SDSC's historical track record of excellence in advanced cyberinfrastructure for the national research community.

### Funding and Grants

In what is now the most competitive landscape for federal funding for scientific research, SDSC's success rate has held steady at about 45%, compared with a national average of about 18% for computer science and engineering proposals at the NSF. As of June 30, 2017, SDSC had an active portfolio of more than 100 contracts and grants totaling about \$127 million; for FY2016-17, some \$27 million in committed funds are being expended against these awards. Based on submitted proposals now under review, the projection for FY2017-18 is for an additional \$15 million. This is in line with SDSC's previous annual records, notwithstanding the ups and downs inherent in the competitive proposal process.

At the highest levels, an increasingly diverse funding portfolio has been a key to SDSC's success. This includes a rich blend of federal funding, which in turn enables and encourages support at the local and state levels, as well as among

(SDSC's) commitment to fulfilling its national advanced cyberinfrastructure mission provides significant benefits to UC San Diego, the UC system, and our industry partners.

our industry partners. One recent example of this diversification is an agreement with the prestigious Simons Foundation's Flatiron Institute to use SDSC's *Gordon* supercomputer for ongoing research in astrophysics, biology, condensed matter physics, materials science, and other domains, following the system's five-year tenure as a key NSF computing resource. We consider this partnership to be a solid testimony regarding *Gordon's* data-intensive capabilities and its myriad contributions to advancing scientific discovery.

I am confident that in reading through this Annual Report, you will appreciate the diversity of SDSC's interests and accomplishments for the good of both science and society. I ask you to keep in mind that while this report details SDSC's impact and influence at the local, state, and national levels, our standing as a national center of high-performance computing and data management is foundational to how we add value to UC San Diego, the UC system, and our industry partners.

*Michael L. Norman*  
SDSC Director

## NEW APPOINTMENT



In early June 2017, SDSC appointed Christopher Irving manager of the Center's High-Performance Computing (HPC) systems. Irving, who joined SDSC in late 2011, has been an HPC systems engineer at SDSC for the past five years. Irving, along with other researchers and HPC experts at SDSC, is also responsible for learning

and testing new technologies for use on SDSC's HPC systems.

Before joining SDSC, Irving worked at The Scripps Research Institute as a systems administrator in the Automated Mo-

lecular Imaging group, a cryo-electron (cryoEM) microscopy laboratory. Prior to that, he was with the Massachusetts Institute of Technology as a systems and network administrator in MIT's Department of Brain and Cognitive Sciences and The Picower Institute for Learning and Memory. Irving holds a B.A. in Computer Science and History from UC Berkeley.

"Christopher has been involved in many facets of deploying and supporting both our *Gordon* and *Comet* supercomputers, so this appointment is a natural fit for all of us," said Amit Majumdar, director of SDSC's Data Enabled Scientific Computing division. "He also has been coordinating closely with our User Services Group in his previous role, so he's now officially overseeing SDSC's high level of providing HPC and data resources for our broad user community."

# INNOVATION IN DATA-ENABLED SCIENTIFIC COMPUTING & BRAIN RESEARCH

## MEET: **AMIT MAJUMDAR**



**A**mit Majumdar, director of SDSC's Data-Enabled Scientific Computing division and an associate professor in the Department of Radiation Medicine and Applied Sciences at UC San Diego, was recently named SDSC's fourth "Pi Person of the Year". Named after the  $\pi$  symbol, this award recognizes researchers who have one 'leg' in a science domain – in this instance the neurosciences – and the other in cyberinfrastructure technologies such as data-enabled high-performance computing.

Majumdar participated in a recently published study that used SDSC's *Comet* supercomputer and Neuroscience Gateway to create a path to developing realistic "biomimetic neuroprosthetics" – or brain implants that replicate neural circuits and their function – that one day could replace lost or damaged brain cells or tissue from tumors, strokes, or other diseases. Majumdar received his B.S. from Jadavpur University in Calcutta, India, in 1985; his M.S. from Idaho State University in 1988; and his Ph.D. from the University of Michigan in 1996. He joined SDSC in 1997.

**Q** Can you describe your involvement in the neuroprosthetics study recently published in the *IBM Journal of Research and Development*?

**A** We collaborated with researchers at the State University of New York (SUNY) Downstate Medical Center by helping to implement a novel computer algorithm on SDSC's *Comet* supercomputer to mimic a neural circuitry that resembles how an unimpaired brain controls limb movement, in this case to direct a realistic prosthetic arm. *Comet* provided the capability to quickly simulate and evaluate thousands of possible models, while the Neuroscience Gateway (NSG) based at SDSC provided a portal to these computational resources. (Majumdar is the principal investigator of the NSG project. For more details about the study see the link at the end of this article.)

**Q** The Neuroscience Gateway has been mentioned as a potential resource for those involved with the national Human Brain Project. Can you describe how this gateway is being used by the neuroscience community today and your expectations for the future?

**A** NSG helps neuroscientists address computationally demanding problems such as simulation of large-scale network models and extraction of connectivity information from brain imaging data. It has become an important tool for developers of computational neuroscience software from several major universities and research institutions for scientists to share their research products with the broader neuroscience community. We already have a user community of over 450, and their yearly usage of supercomputing time is approaching 10 million core-hours across HPC systems at SDSC, the Texas Advanced Computing Center, the Pittsburgh Supercomputing Center, and Indiana University. As a part of NSG we are also getting involved with various neuroscience training programs funded by the National Science Foundation (NSF) and the National Institutes of Health.

**Q** How is the Data-Enabled Scientific Computing group structured and what capabilities does it offer users?

**A** DESC is responsible for the High-Performance Computing Systems group, User Services group, Scientific Computing Applications group, Scientific Visualization group, and the Advanced Technology Lab. SDSC's involvement in the NSF's Extreme Science and Engineering Discovery Environment, or XSEDE program, is coordinated from the DESC division. DESC staff members have degrees in domain sciences such as chemistry, computer science, physics, applied mechanics, astrophysics, bioinformatics, and various branches of engineering. All staff members have expertise in high-performance computing, data-intensive computing, and scientific software. They design and maintain HPC systems and provide user support for our supercomputers, and work on funded research projects involving various scientific applications, including big data problems

**Q** Can you tell me more about the new Advanced Technology Lab?

**A** The ATL, in operation just over a year, is focused on new technologies, be it hardware or software or combinations thereof and technologies that can impact application performance on HPC and data resources. ATL researchers focus on gaining access to newest technologies, in many cases before they are generally available or while under development. One of the ATL's recent research projects includes looking into data movement at a lower level within processor

architectures, which is becoming a determining factor in processor architectures from the performance and energy aspects. Another project involved performance characterization of biosciences applications on multi-core processors for widely used bioinformatics and cryo-electron microscopy data analysis software.

**Q** Regarding your appointment in UC San Diego's Department of Radiation Medicine, how can HPC play a role in radiation therapy?

**A** We work with radiation oncologists to see how the computationally-intensive parts of patient treatment plans can be made more efficient using high-performance computing, GPUs, etc. Computationally-intensive parts include image analysis, dose calculation, and optimization of radiation dose delivery. Separately, analysis of patient data, which is a data science problem in of itself, can also provide insight into effectiveness of treatments and long-term impacts.

**Q** Today, data-enabled science and life science research are dependent upon computationally-based research – in fact they wouldn't be able to flourish without HPC. How are these areas intertwined and what do you envision in the next five or 10 years?

**A** Analysis of data has become a science in itself. Computer scientists, statisticians, mathematicians, and computational scientists are getting involved with domain scientists to understand how data science algorithms such as in machine learning can be applied to solve new and different kinds of problems. In neuroscience, and as a result of advanced experimental techniques, larger and complex data sets are being produced. Data scientists are now working with neuroscientists, bringing their data analytics knowledge. All of this will have an impact, for example, on new HPC architectures suitable for machine learning. HPC systems will have to cater to both computational scientists and data scientists by designing comprehensive HPC machines with the architecture and software stack to serve both communities.



Learn more about the NSG project by scanning the QR code on the left or by visiting <https://goo.gl/bPS3kl>





# GATEWAYS TO DISCOVERY: SDSC's COMPUTATIONAL RESOURCES

SDSC's computational, storage, and networking resources – along with the human expertise to operate and support them – form an advanced cyberinfrastructure used to accelerate scientific discovery at the local, state, national, and even global levels.

Advanced but also user-friendly resources such as SDSC's petascale-level *Comet* supercomputer underscore a strong need for systems that serve a diverse group of research domains, as well as the majority of scientists who have modest-scale computational needs. *Comet* continues to be one of the most widely used supercomputers in the National Science Foundation's XSEDE (Extreme Science and Engineering Discovery Environment) program, which provides academic researchers with an advanced collection of integrated digital resources and services.

"We as a nation require a next-generation cyberinfrastructure that supports effective and efficient collaborations that will drive us toward grand challenge discoveries," said SDSC Director Michael Norman, who is also the principal investigator for the *Comet* program, the result of an NSF grant now totaling almost \$25 million.

(right) Inside SDSC's 19,000 square foot datacenter.





*Comet, housed in SDSC's main data center, is configured to serve a wide range of researchers in both traditional and non-traditional science domains.*

# COMET

## HPC FOR THE 99 PERCENT

Within its first 18 months of operation, SDSC's petascale-level *Comet* supercomputer soared past its project goal of serving 10,000 unique users across a diverse range of science disciplines, from astrophysics to redrawing the "Tree of Life."

In fact, more than 15,000 users have used *Comet* to run science gateways jobs alone since the system went into production in mid-2015. A science gateway is a community-developed set of tools, applications, and data services and collections that are integrated through a web-based portal or suite of applications. Another 2,600 users have accessed the system via traditional login.

The 10,000-user target was established by SDSC as part of its cooperative agreement with the NSF, which awarded funding for *Comet* in late 2013. Since entering operations, *Comet* has provided more than 600 million core-hours of computing across a wide range of science disciplines.

"*Comet* was designed to meet the needs of what is often referred to as the 'long tail' of science – the idea that the large number of modestly-sized computationally-based research projects represent, in aggregate, a tremendous amount of research that can yield scientific advances and discovery," said Norman.

While *Comet* is capable of an overall peak performance of two petaflops – that's two quadrillion calculations per second –

its allocation and operational policies are geared toward rapid access, quick turnaround, and an overall focus on scientific productivity. "*Comet* represents what we like to call HPC for the 99 percent," said Norman. "It's about providing high-performance computing to a much larger research community, while meeting the needs of underserved researchers in domains which have not traditionally relied on supercomputers to help solve problems."

*Comet* has become a workhorse system for a large fraction of computational research provided through XSEDE. Its unique architecture and operational policies that target the long tail of science provide a blueprint for how such systems can be designed and supported. *Comet* lowers barriers to HPC through the use of trial accounts, allocation policies that favor modest-scale computing and gateways, and system management policies that ensure responsiveness and throughput. With the growth of these new computing modalities and approaches SDSC expects that systems such as *Comet* will continue to be essential to NSF's computing portfolio. Shortly after the close of the fiscal year ended June 30, 2017, *Comet* ran its ten millionth job, another milestone that demonstrates the tremendous amount of productive computing the system has run.

## GPU COUNT DOUBLED

SDSC recently doubled the number of graphic processing units (GPUs) on *Comet* in direct response to growing demand for GPU computing among a wide range of research domains. The expansion makes *Comet* the largest provider of GPU resources available to the NSF's XSEDE program, a national partnership of institutions that provides academic researchers with the most advanced collection of digital resources and services in the world.

Once used primarily for video game display graphics, today's much more powerful GPUs have been developed that have more accuracy, speed, and accessible memory for more scientific applications that range from phylogenetics and molecular dynamics to creating some of the most detailed seismic simulations ever made to better predict ground motions to save lives and minimize property damage.

Under the supplemental NSF award, valued at just over \$900,000, SDSC expanded *Comet* with the addition of 36 GPU nodes, each with four NVIDIA P100s, for a total of 144 GPUs. This doubles the number of GPUs from the current 144 to 288.

Applications include but are not limited to GPU-memory management systems such as VAST, analysis of data from large scientific instruments, and molecular dynamics software packages such as AMBER, LAMMPS, and BEAST – the latter used extensively by SDSC's Cyberinfrastructure for Phylogenetic Research (CIPRES) science gateway, which receives the majority of its computing resources from *Comet*.

"This expansion will help the XSEDE organization meet increased demand for GPU resources from these areas, as well as prepare for research in new areas, such as machine learning, which has become increasingly important for a wide range of research in areas including image processing, bioinformatics, linguistics, and others," said SDSC Director of Scientific Applications and *Comet* Co-PI Bob Sinkovits.

The amended NSF award number for *Comet*, including the GPU additions, is FAIN 1341698. The award is estimated to run until March 30, 2020.



Watch a video about *Comet* by scanning the QR code on the left or by visiting <https://youtu.be/2rX9antVMRw>

## Comet: A Science Gateways Leader

Web-based science gateways make it possible to run the available applications on supercomputers such as *Comet* so results come quickly, even with large data sets. Moreover, browser access offered by gateways allows scientists to focus on their research without having to learn the details of how supercomputers work or how to access and organize the data needed.

Some 32 science gateways are available via XSEDE's resources, covering a wide range of domains. *Comet* has been one of the most widely used supercomputers for such gateways; since the start of the XSEDE project in 2011, SDSC alone has delivered more than 75 percent of all gateway cycles.

In mid-2016, a collaborative team led by SDSC was awarded a five-year \$15 million NSF grant to establish a Science Gateways Community Institute (SGCI) to accelerate the development and application of highly functional, sustainable science gateways that address the needs of researchers across the full spectrum of NSF directorates. (Read more about SGCI on page 40)

In late 2016, a new science gateway called I-TASSER (Iterative Threading ASSEmbly Refinement), developed by researchers at the Zhang Lab at the University of Michigan's Medical School, began accepting users. I-TASSER is a hierarchical approach to protein structure and function prediction. Structural templates are first identified from the Protein Data Bank. The new gateway is only available on *Comet*.

## SDSC Shares Awards for HPCwire's 'Top Supercomputing Achievement'

SDSC was a recipient of two HPCwire 'Top Supercomputing Achievement' awards for 2016, recognizing the use of *Comet* to help verify Albert Einstein's theory of gravitational waves as part of a landmark discovery. In February 2016, the NSF announced that for the first time, scientists detected gravitational waves in the universe as hypothesized by Einstein about 100 years ago. On September 14, 2015, scientists at the NSF-funded Laser Interferometer Gravitational-Wave Observatory (LIGO) detected gravitational waves using both LIGO detectors. LIGO used *Comet*'s Virtual Cluster interface, provided via the Open Science Grid (OSG), for the analysis of the data.

The awards, won in both the online publication's annual Readers' Choice and Editors' Choice categories, also recognized OSG, the NSF's Extreme Science and Engineering Discovery Environment (XSEDE), the Holland Computing Center at University of Nebraska-Lincoln (UNL), the National Center for Supercomputing Applications (NCSA), and the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for their participation in verifying the existence of gravitational waves.

The awards were presented at the 2016 International Conference for High Performance Computing, Networking, Storage and Analysis (SC16), in Salt Lake City, Utah. **See pages 19 and 22 for research that won SDSC and its partners two key HPCwire Awards for 2017.**



# 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 to use massive amounts of flash-based memory. That made it many times faster than conventional HPC systems, while having enough bandwidth to help researchers sift through tremendous amounts of data. The result of a \$20 million NSF grant, *Gordon* remained an NSF resource until the end of March 2017 following two extensions of service.

During its NSF tenure, *Gordon* supported research and education by more than 2,000 command-line users and over 7,000 gateway users, primarily through resource allocations from XSEDE. One of *Gordon's* most data-intensive tasks was to rapidly process 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.

## Gordon Gets a New Lease on Life

*Gordon* continues to advance scientific discovery under a new partnership with the Simons Foundation's Flatiron Institute in New York. Under a two-year agreement, the majority of the data-intensive supercomputer is being used by the Institute for ongoing research in astrophysics, biology, condensed matter physics, materials science, and other domains. SDSC has retained a smaller portion of *Gordon* for use by other organizations including UC San Diego's Center for Astrophysics & Space Sciences (CASS), as well as SDSC's OpenTopography project and various projects within the Center for Applied Internet Data Analysis (CAIDA), which is based at SDSC.

"We are delighted that the Simons Foundation has given *Gordon* a new lease on life after five years of service as a highly sought-after XSEDE resource," said SDSC Director Michael Norman, who also served as the principal investigator for *Gordon*. "We welcome the Foundation as a new partner and consider this to be a solid testimony regarding *Gordon's* data-intensive capabilities and its myriad contributions to advancing scientific discovery."

Specifically, *Gordon* is being used in conjunction with the Simons Observatory, a five-year \$40 million project awarded by the Foundation in May 2016 to a consortium of universities led by UC San Diego, UC Berkeley, Princeton University, and the University of Pennsylvania. In the Simons Observatory, new telescopes are joining the existing POLARBEAR/Simons Array and Atacama Cosmology Telescopes to produce an order of magnitude more data than the current POLARBEAR experiment.



Learn more about SDSC's HPC Systems by scanning the QR code on the left or by visiting <https://goo.gl/McdvqA>

# TSCC COMPUTING “CONDO”

## AFFORDABLE COMPUTING FOR CAMPUS AND CORPORATE USERS

One of SDSC’s major contributions to research computing at UC San Diego has been the *Triton Shared Computing Cluster (TSCC)*, a “condo computing” program established in 2013 that has continued to grow in popularity among campus and external users during the FY2016-17 period. In the four years since its launch, *TSCC* has grown to include 30 labs/groups with more than 200 researcher-owned compute nodes, plus an additional 75 common nodes available to anyone on campus through a pay-as-you-go recharge model. The latter is popular with individual researchers with occasional or temporary computing needs, students, and classes covering parallel architecture.

In early 2017, SDSC was awarded an NSF grant to upgrade *TSCC* to deliver targeted capabilities for bioinformatics analyses. The grant, valued at almost half-a-million dollars and slated to run through January 2018, is part of the NSF’s Campus Cyberinfrastructure (CC\*) program, which invests in coordinated campus-level cyberinfrastructure (CI) components of data, networking, computing infrastructure, capabilities, and integrated services. A key objective of the project is to leverage new technologies to provide accelerated computing capacity so that researchers can conduct high throughput analyses of many whole genomes, while also having the ability to conduct quick turnaround, single-genome analyses. The latter capability could be particularly useful for precision medicine and emerging clinical applications of genomics.

### ‘BIOBURST’

Under the NSF award, SDSC is implementing a separately scheduled partition of *TSCC* with technology designed to address key areas of bioinformatics computing including genomics, transcriptomics, and immune receptor repertoire analysis. Called ‘*BioBurst*’, the system will incorporate the following major components:

- An input/output (I/O) accelerator appliance with high-performance, non-volatile memory and software designed to improve network throughput by alleviating the small-block/small-file I/O problem characteristic of many bioinformatics codes;
- A field programmable gate array (FPGA)-based computational accelerator system that has been demonstrated to perform de-multiplexing, read mapping, and variant calling of complete human genomes in less than one-hour timescales;
- Several hundred new commodity computing cores, which will access the I/O accelerator and provide a separately scheduled resource for running bioinformatics applications;
- Integration with a large-scale, Lustre parallel file system which supports streaming I/O and has the capacity to stage large amounts of data associated with many bioinformatics studies; and
- Customization of the job scheduler to accommodate bioinformatics workflows, which can consist of hundreds to thousands of jobs submitted by a single user at one time.





## COLO FACILITY

### **SERVING THE LOCAL RESEARCH COMMUNITY**

SDSC offers rack colocation (colo) services to the local research community. Its 19,000 square-foot climate-controlled and secure data center is fully equipped with 13 megawatts of power, multi- 10-gigabit network connectivity, and a 24/7 operations staff. UC San Diego colocation is supported via a program aimed at saving campus funds by housing equipment in an energy-efficient, shared facility. UC partners and research collaborators are also able to make use of the facility.

The colo facility houses computing and networking equipment for 34 campus departments, every division and school, as well as local partners that include Rady Children's Hospital, J. Craig Venter Institute, the Simons Foundation, The Scripps Research Institute, and the Sanford-Burnham Medical Research Institute.

SDSC's colo facility has resulted in more than \$2 million a year in energy savings, while streamlining and improving the management of hundreds of campus systems. It also has facilitated numerous collaborations with and between UC San Diego researchers, who can more easily share data, integrate with SDSC compute resources, and access national high-performance research networks. The facility is especially well-suited to installations that must demonstrate regulatory compliance, as well as require high-speed networking, monitoring, high levels of uptime, or backup power.

# DATA OASIS

## AMONG ACADEMIA'S FASTEST PARALLEL FILE SYSTEMS

SDSC's *Data Oasis* is a Lustre-based parallel file storage system linked to *Comet*, *Gordon*, and *TSCC*. As a critical component of SDSC's big data initiatives, *Data Oasis* currently has about 12 petabytes (PB) of capacity and speeds of up to 200 gigabytes (GB) per second to handle just about any data-intensive project. *Data Oasis* ranks among the fastest parallel file systems in the academic community. Its sustained speeds mean researchers could retrieve or store 240 terabytes (TB) of data—the equivalent of *Comet*'s entire DRAM memory—in about 20 minutes, significantly reducing time needed for retrieving, analyzing, storing, or sharing extremely large datasets. In short, *Data Oasis* allows researchers to analyze data at a much faster rate than most other systems, which in turn helps extract knowledge and discovery from these datasets. In early 2015, *Data Oasis* underwent significant upgrades, including ZFS, a combined file system originally designed by Sun Microsystems and mated in a new hardware server configuration.

# SDSC CLOUD

## LARGE-SCALE ACADEMIC DEPLOYMENT OF CLOUD STORAGE AND COMPUTE

The SDSC IT Services team administers a large-scale storage and compute cloud. UC San Diego campus users, members of the UC community, and UC affiliates are eligible to join the hundreds of users who already benefit from the multi-petabyte, OpenStack's Swift object store. *SDSC Cloud* is the perfect storage choice for researchers with fixed budgets because unlike other cloud providers, *SDSC Cloud* boasts a simplified recharge plan that eliminates secondary fees such as bandwidth costs, charges assessed per request, and regional migration fees. *SDSC Cloud* also includes an elastic compute facility, based on OpenStack Nova and Ceph. This comprehensive cloud environment provides researchers with a testbed and development environment for developing cloud-based services. It is especially attuned to data sharing, data commons and data analytics services. *SDSC Cloud* is one of the platforms that underpins the National Data Service suite of offerings. Beginning in July 2017, researchers who made use of *SDSC Cloud* using sponsored research funds no longer paid IDC. This is just one way that SDSC and UC San Diego are working together to make the campus a research-friendly environment.



# NETWORKING & CONNECTIVITY

SDSC has helped lay the groundwork and provide expertise in implementing networks that allow fast and unrestricted flow of information between systems and researchers, both on and off the UC San Diego campus.

## PACIFIC RESEARCH PLATFORM

### ADVANCING COLLABORATION UP AND DOWN THE WEST COAST

From biomedical data to particle physics, researchers depend heavily on high-speed access to large datasets, scientific instruments, and computing resources. The NSF funded a \$5 million, five-year award for UC San Diego and UC Berkeley to establish a high-capacity, data-centric “freeway system” that will give participating universities and other research institutions the ability to move data about 1,000 times faster than what’s possible on today’s inter-campus shared internet.

SDSC is an anchor participant in this project, called the Pacific Research Platform, or PRP. The PRP links numerous research universities on the West Coast – including the 10 UC campuses, San Diego State University (SDSU), Caltech, USC, Stanford, and University of Washington – via the Corporation for Education Network Initiatives in California (CENIC)/Pacific Wave’s 100G infrastructure. The initiative also extends to Lawrence Berkeley National Laboratory, the National Energy Research Scientific Computing Center, NASA Ames, and the National Center for Atmospheric Research and several campuses outside of California, including international campuses. Partner networks include the Pacific Northwest Gigapop, the Energy Sciences Network, NASA Research and Engineering Network, the Metropolitan Research & Education Network, StarLight, and the Front Range Gigapop – with a long-term goal of engaging other research and education networks in the U.S. and abroad.

Other PRP/CENIC goals include the integration of radio and wired networks over a wide area supporting access to regional sensors and devices, as well as multi-campus resident data services. Goals include the design to support resilience of communications across San Diego and Orange County should any of the campus or mountain top relay points become unreachable. The testbed for this multi-homed radio/wired network is a work in progress as researchers evaluate the needed network infrastructure routing and announcement re-configurations required to support such resilience. Participants include SDSC/UC San Diego, UC Irvine, SDSU, and CENIC. The procedures being developed to support such a regional radio network expansion will provide CENIC with a blueprint for other similar radio based network expansions around the state of California. Read more about the PRP on page 33.

## PRISM@UCSD

### THE HOV LANE FOR BROADBAND 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. This network extends the ‘Science DMZ’ concept into a Distributed Science DMZ. Called Prism@UCSD and backed by a \$500,000 NSF grant, researchers with SDSC and the campus’ California Institute for Telecommunications and Information Technology (Calit2) 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. “One can think of Prism as the HOV lane, whereas our very capable campus network represents the other lanes on the freeway,” said Philip Papadopoulos, principal investigator on the Prism@UCSD project and SDSC’s chief technology officer.

## CHERUB

### CONNECTING TO THE INFORMATION SUPERHIGHWAY

Called CHERuB for Configurable, High-speed, Extensible Research Bandwidth, this project is the result of a second \$500,000 NSF grant awarded to SDSC and UC San Diego’s Administrative Computing and Telecommunications (ACT) organization to connect the campus’s Science DMZ (Prism) and SDSC resources to high-bandwidth national research networks to advance a new range of data-driven research. With 100 gigabit-per-second external connectivity, CHERuB supports multi-institutional data transit over networks such as the Internet2’s Advanced Layer 2 Service (AL2S), the Department of Energy’s ESnet, Pacific Wave and CENIC as well as a joint project among 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.” CHERuB also supports multiple HPC projects and programs SDSC and UC San Diego such as *Comet*, *Gordon*, *TSCC*, the Physics LHC project, and PRP. Research domains that can benefit from CHERuB include cosmology, atmospheric sciences, electron microscopy, genomic sequencing, oceanography, high-energy physics, and telemedicine – all of which encompass data-intensive research.

“CHERuB allows us to extend capabilities of our Distributed Campus Science DMZ (Prism@UCSD) to other institutions and allows for new forms of collaboration between institutions,” says SDSC Network Architect and CHERuB co-PI Thomas Hutton.





SCIENCE  
HIGHLIGHTS



# BIODIVERSITY

REVEALING AN UNDERGROUND WORLD  
OF STUNNING DIVERSITY

**Stunning diversity, visualized.** All the known major bacterial groups are represented by wedges in this circular “tree of life.” The bigger wedges are more diverse groups. Green wedges are groups that have not been genomically sampled at the Rifle, CO research site — everything else has. Black wedges are previously identified bacteria groups that have also been found at Rifle. Purple wedges are groups discovered at Rifle and announced last year. Red wedges are new groups discovered in this study. Colored dots represent important metabolic processes the new groups help mediate. Credit: Banfield Group

One of the most detailed genomic studies of any ecosystem to date reveals an underground world of stunning microbial diversity, resulting in the addition of dozens of new branches to the ‘tree of life.’ These findings shed light on one of Earth’s most important and least understood realms of life as the subterranean world hosts up to one-fifth of all biomass but still remains a mystery.

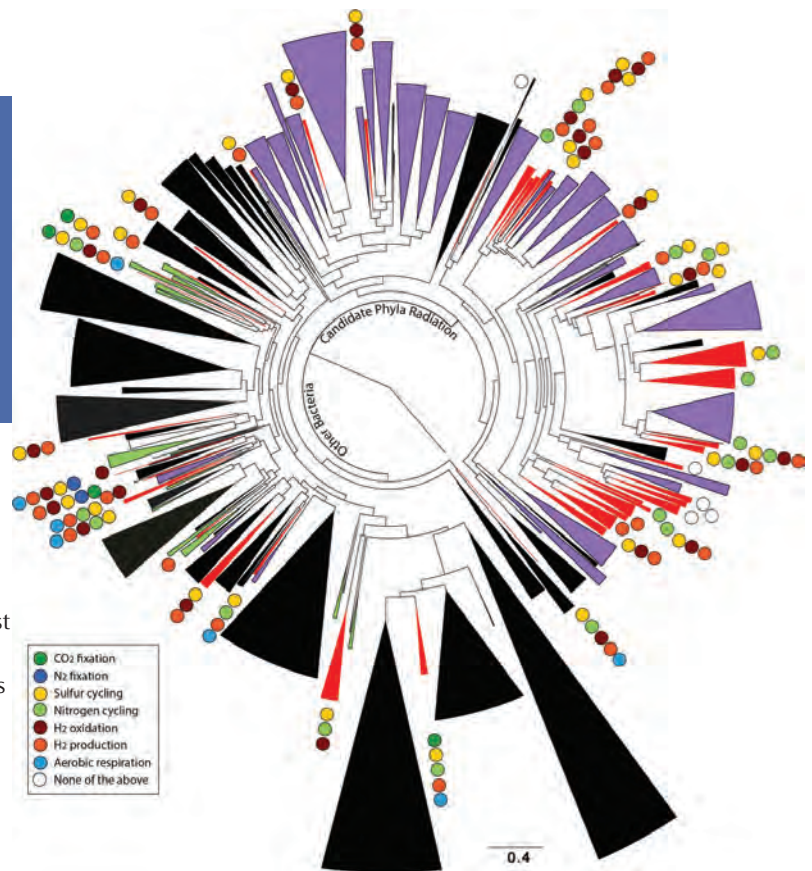
This bacterial bonanza was uncovered by scientists who reconstructed the genomes of more than 2,500 microbes from sediment and groundwater samples collected at an aquifer in Colorado. The study, reported in the October 24, 2016 online edition of *Nature Communications*, was led by researchers from the Department of Energy’s (DoE) Lawrence Berkeley National Laboratory (Berkeley Lab) and UC Berkeley.

Scientists netted genomes from 80 percent of all known bacterial phyla, a remarkable degree of biological diversity at one location. They also discovered 47 new phylum-level bacterial groups, naming many of them after influential microbiologists and other scientists. They also learned new insights about how microbial communities work together to drive processes that are critical to the planet’s climate and life everywhere, such as the carbon and nitrogen cycles.

“We didn’t expect to find this incredible microbial diversity,” said Jill Banfield, a senior faculty scientist in Berkeley Lab’s Climate & Ecosystem Sciences Division and a UC Berkeley professor in the departments of Earth and Planetary Science, and Environmental Science, Policy, and Management. “Then again, we know little about the roles of subsurface microbes in biogeochemical processes, and more broadly, we don’t really know what’s down there.”

To better understand what subsurface microbes are up to, the researchers’ approach was to access their entire genomes. That enabled them to discover a greater interdependency among microbes than seen previously.

The scientists sent soil and water samples from these experiments to the Joint Genome Institute for terabase-scale




metagenomic sequencing. This high-throughput method isolates and purifies DNA from environmental samples, and then sequences one trillion base pairs of DNA at a time. Next, the scientists used bioinformatics tools developed in Banfield’s lab along with those from SDSC’s CyberInfrastructure for Phylogenetics REsearch, or CIPRES Gateway, and analyses were conducted with the aid of the SDSC’s *Comet* supercomputer.

“The CIPRES Science Gateway and the *Comet* supercomputer were instrumental to our work,” Banfield said. “Considering the unprecedented size of our sequence datasets, we were unable to complete any runs for inferring trees on other servers.”

The scientists’ approach has redrawn the tree of life. Between the 47 new bacterial groups reported in this work, and 35 new groups published the previous year, Banfield’s team has doubled the number of known bacterial groups. Another big outcome is a deeper understanding of the roles subsurface microbes play in globally important carbon, hydrogen, nitrogen, and sulfur cycles. This information will help to better represent these cycles in predictive models such as climate simulations.

Read more about *Comet* on page 8 and more about the CIPRES Gateway on page 41.



# ARTIFICIAL INTELLIGENCE QUALITY OF LIFE

REPLICATING BRAIN CIRCUITRY TO  
DIRECT A REALISTIC PROSTHETIC ARM



SDSC received the HPCwire "Readers' Choice - Best Use of AI" award for using SDSC's *Comet* supercomputer to develop realistic "biomimetic neuroprosthetics" by replicating brain circuitry to direct a realistic prosthetic arm.

## NOVEL ALGORITHM MAY SPEED SILICON IMPLANTS TO CORRECT BRAIN DAMAGE

By applying a novel computer algorithm to mimic how the brain learns, a team of researchers – with the aid of SDSC’s *Comet* supercomputer and the Center’s Neuroscience Gateway – has identified and replicated neural circuitry that resembles the way an unimpaired brain controls limb movement.

The research, published in the March-May 2017 issue of the *IBM Journal of Research and Development*, lays the groundwork to develop realistic “biomimetic neuroprosthetics” – brain implants that replicate brain circuits and their function – that one day could replace lost or damaged brain cells or tissue from tumors, stroke, or other diseases.

“In patients with motor paralysis, the biomimetic neuroprosthetic could be used to replace the deteriorated motor cortex where it could interact directly with healthy brain pre-motor regions, and send commands and receive feedback via the spinal cord to a prosthetic arm,” said W.W. Lytton, a professor of physiology and pharmacology at State University of New York (SUNY) Downstate Medical Center in Brooklyn, N.Y., and the study’s principal investigator.

This scenario, portrayed in the IBM paper titled “Evolutionary algorithm optimization of biological learning parameters in a biomimetic neuroprosthesis”, required high-performance computing and expertise to simulate and evaluate potential computer models in an automated way, along with the Neuroscience Gateway (NSG) based at SDSC, which provided an entrance to these resources.

“The increasing complexity of the virtual arm, which included many realistic biomechanical processes, and the more challenging dynamics of the neural system, called for more sophisticated methods and highly-parallel computing in a

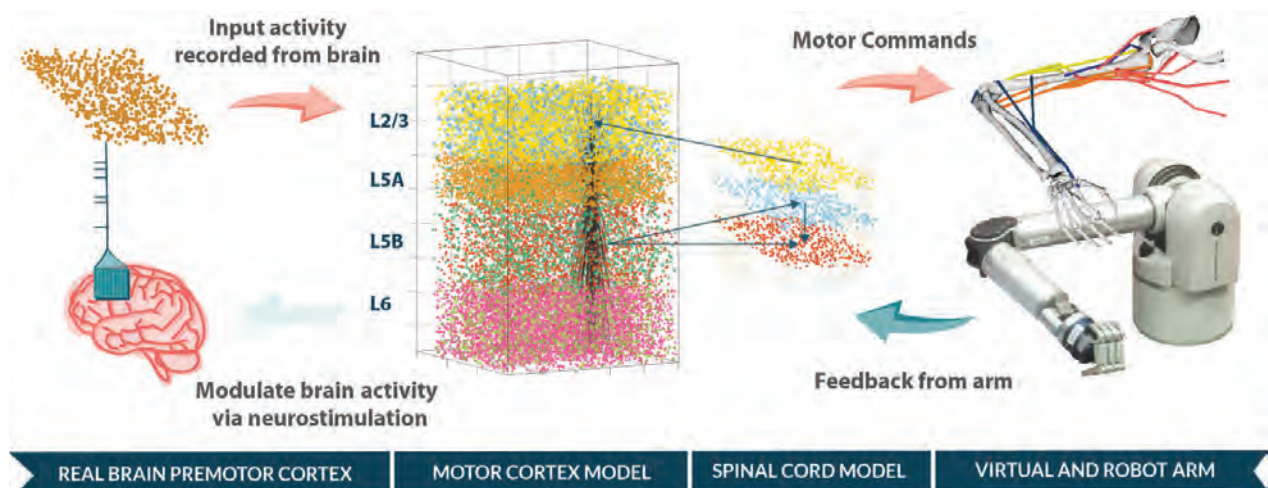
system such as *Comet* to tackle thousands of model possibilities to come up with ever more realistic prostheses,” said Amit Majumdar, director of SDSC’s Data Enabled Scientific Computing division, principal investigator of the NSG, and co-author of the IBM Journal paper. Subhashini Sivagnanam, a senior scientific computing specialist at SDSC and co-PI of the NSG project, is a co-author of the IBM paper.


## FUSING COMPUTATIONAL AND BIOLOGICAL PRINCIPLES

Over the past decade or so, researchers have been trying to fuse computational and biological principles to create realistic computer models that would form the basis for silicon-based neural circuits or implants that would replace damaged brain tissue. In this emerging field, a primary goal has been the decoding of electrical signals recorded from the brain to move, for example, a prosthetic arm. In scenarios once considered science fiction, techniques that encode neural signals from a prosthetic virtual arm to the brain are now allowing users to feel what they are touching.

That said, researchers now recognize that controlling even more realistic and complex systems, including prosthetic limbs involving larger numbers of bones, joints and muscles, require computer models that more closely resemble real brain circuitry. Future studies will focus on developing even more realistic models of the primary motor cortex microcircuits to help understand and decipher the neural code, or how information is encoded and transmitted in the brain.

**Overview of biomimetic neuroprosthetic system.** Left to right: Information about what target to reach can be gathered from electrodes in the brain. This modulates ongoing activity in the biomimetic cortical and spinal cord models which then drives the virtual arm, which is then mirrored by the robot arm. Right to left: haptic feedback (touch sensation) could then be delivered back in the other direction so that the user could feel what is being touched. *Reproduced with permission from Dura-Bernal et al. 2017 (IBM Journal of Research and Development)*





# MATERIALS ENGINEERING

## DEVELOPING PLASTIC FABRIC THAT COOLS THE SKIN

Stanford University researchers, with the aid of SDSC's *Comet* supercomputer, have engineered a low-cost plastic material that could become the basis for clothing that cools the wearer, reducing the need for energy-consuming air conditioning.

Describing their work in the September 2, 2016 issue of *Science*, the researchers suggested that this new family of fabrics could become the basis for garments that keep people cool in hot climates without air conditioning. "If you can cool the person rather than the building where they work or live, that will save energy," said Yi Cui, an associate professor of materials science and engineering at Stanford University and of photon science at SLAC National Accelerator Laboratory, and the study's principal investigator.

The new material works by allowing the body to discharge heat in two ways that would make the wearer feel nearly 4 degrees Fahrenheit cooler than if they wore cotton clothing.

The material cools by letting perspiration evaporate through it, something ordinary fabrics already do. But the Stanford material provides a second, revolutionary cooling mechanism: allowing heat that the body emits as infrared radiation to pass through the plastic textile.

All objects, including our bodies, throw off heat in the form of infrared radiation, an invisible and benign wavelength of light. Blankets warm us by trapping infrared heat emissions close to the body. This thermal radiation escaping from our bodies is what makes us visible in the dark through night-vision goggles.

“Forty to 60 percent of our body heat is dissipated as infrared radiation when we are sitting in an office,” said Shanhui Fan, co-author of the study and a professor of electrical engineering who specializes in photonics, which is the study of visible and invisible light. “But until now there has been little or no research on designing the thermal radiation characteristics of textiles.”

## SUPER-POWERED KITCHEN WRAP

The research blended computer simulations, nanotechnology, photonics and chemistry to give polyethylene – the clear, clingy plastic we use as kitchen wrap – a number of characteristics desirable in clothing material: it allows thermal radiation, air and water vapor to pass right through and is opaque to visible light. The easiest attribute was allowing infrared radiation to pass through the material, because this is a characteristic of ordinary polyethylene food wrap. Of course, kitchen plastic is impervious to water and is see-through as well, rendering it useless as clothing.

As an initial step, researchers created computer models that captured the optical properties of nanoporous polyethylene. Simulations were conducted on a local compute cluster at

Stanford, in addition to SDSC’s *Comet* supercomputer and *Stampede* at the Texas Advanced Computing Center at the University of Texas in Austin. The resulting models encompassed a wide optical wavelength range, from the visible to infrared.

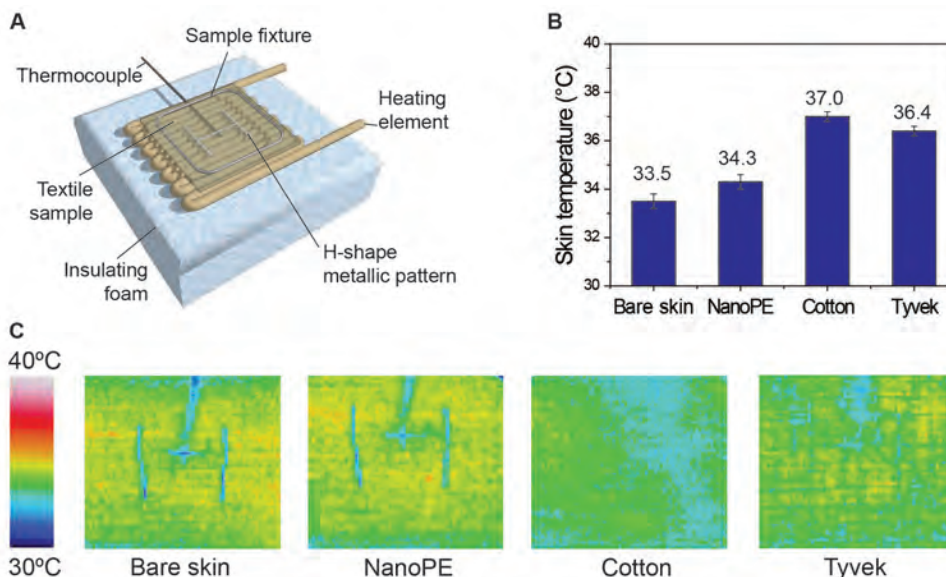
“Solving for electromagnetic wave propagation in large 3D structures is very computationally demanding, and can only be done on high-performance computers,” said Fan. “Otherwise it would take too much time. The large shared memory in the *Comet* cluster was quite beneficial for the code that we were using.”

## MORE COLORS, MORE TEXTURES, MORE CLOTH-LIKE

That success gave the researchers a single-sheet material that met their three basic criteria for a cooling fabric. To make this thin material more fabric-like, they created a three-ply version: two sheets of treated polyethylene separated by a cotton mesh for strength and thickness. To test the cooling potential of their three-ply construct versus a cotton fabric of comparable thickness, they placed a small swatch of each material on a surface that was as warm as bare skin and measured how much heat each material trapped.

The researchers are continuing their work on several fronts, including adding more colors, textures, and cloth-like characteristics to their material. Adapting a material already mass produced for the battery industry could make it easier to create products. “We fully expect high-performance computers to be of critical help in our next simulations involving larger scale and more complex structures,” said Alex Song, a Stanford postdoctoral research associate, who carried out the electromagnetic simulations of these fiber structures.

*Based on an article written by Stanford University Science Writer Tom Abate, with additions by Warren Froelich, SDSC External Relations*



Thermal measurement of nanopolyethylene (nanoPE) and various textile samples. (A) experimental setup of textile thermal measurement. The heating element that generates constant heating power is used to simulate human skin, and the “skin temperature” is measured with the thermocouple. Lower skin temperature means a better cooling effect. (B) Thermal measurement of bare skin, nanoPE, cotton, and Tyvek. NanoPE has a much better cooling effect than that of cotton and Tyvek because of its infrared (IR)-transparency. (C) Thermal imaging of bare skin and the three samples. Only nanoPE can reveal the H-shape metallic pattern because of its IR-transparency. Credit: Stanford University



# HUMAN HEALTH LIFE SCIENCES

## NOVEL MOLECULAR DYNAMICS CAPTURES ATOMIC-LEVEL DETAIL OF CRISPR-CAS9 ACTIVITY



SDSC and its partner institutions received the HPCwire “Editors’ Choice - Best Use of HPC in Life Sciences” award for research that identified structural changes activating the gene-splicing technology called CRISPR-Cas9.

Using a novel molecular dynamics method capable of capturing the motion of gyrating proteins at time intervals up to one thousand times greater than previous efforts, a team led by UC San Diego researchers has identified, for the first time, the myriad structural changes that activate and drive CRISPR-Cas9, the innovative gene-splicing technology that’s transforming the field of genetic engineering.

By shedding light on the biophysical details governing the mechanics of CRISPR-Cas9 (clustered regularly interspaced short palindromic repeats) activity, the study provides a fundamental framework for designing a more efficient and accurate genome-splicing technology that doesn’t yield “off-target” DNA breaks currently frustrating the potential of the CRISPR-Cas9 system, particularly for clinical uses.

“Although the CRISPR-Cas9 system is rapidly revolutionizing life sciences toward a facile genome editing technology, structural and mechanistic details underlying its function have remained unknown,” said Giulia Palermo, a postdoctoral scholar with the UC San Diego Department of Pharmacology and lead author of the study, published in the June 26 early edition of the *Proceedings of the National Academy of Sciences (PNAS)*.

“In particular, we want to design a system that doesn’t cause ‘off-target’ effects or non-selective cleavage of DNA sequences, that can now create unwanted collateral damage,” added J. Andrew McCammon, the Joseph E. Mayer Chair of Theoretical Chemistry at UC San Diego, a Howard Hughes Medical Institute Investigator, and the study’s principal investigator.



## CAPTURING PROTEIN DYNAMICS

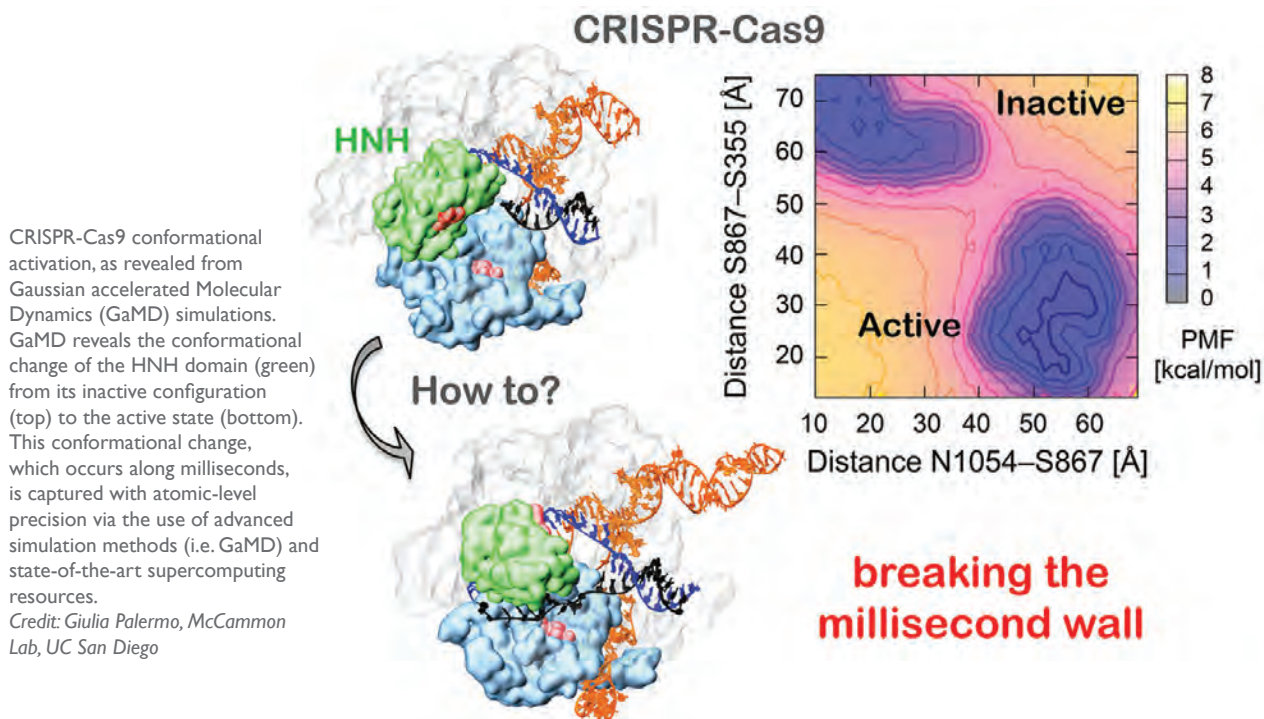
Last year, researchers from the McCammon Lab at UC San Diego used the *Comet* supercomputer at SDSC to perform atomistic molecular dynamics – a method that captures a more complete vision of the myriad shapes and conformations that a target protein molecule may go through – at petascale speeds (one quadrillion arithmetic calculations per second). The resulting simulations, published in the September 8 issue of *ACS Central Science*, identified some of the key factors underlying the complex structural changes taking place during the merger and preparation for cleaving the target DNA strand.

Though explaining some of the fundamental biophysics of the system, these relatively short-lived simulations – conducted upward of multi-microseconds – could not clarify the conformational transitions arising from the protein’s resting state to the catalytically active Cas9 state. To overcome this limitation, the researchers turned to a novel methodology called Gaussian accelerated Molecular Dynamics or GaMD, created by Yinglong Miao in McCammon’s lab, a process that lengthens the observational time-scale from microseconds (millionths of a second) to milliseconds (thousandths of a second). This advance allows researchers to simulate more complex biophysical transitions, including protein folding and ligand binding, while also capturing complex structural transitions.

“The fact that GaMD breaks the millisecond barrier is of great importance, because it gives us the opportunity to observe biological processes that are out of reach of experiments with atomic-level resolution,” said Miao, a co-author of the PNAS study.

Using GaMD, the researchers once again returned primarily to *Comet* – in addition to SDSC’s *Triton Shared Computing Cluster (TSCC)* and *Bridges* at the Pittsburgh Supercomputing Center – to perform CRISPR-Cas9 simulations at the lengthened time-scale. The results provided “on-the-fly” atomic-level detail of conformational changes that take place in the CRISPR-Cas9 system, leading the Cas9 protein from its resting state to its final RNA-bound catalytic state. Until now, structural information about active Cas9 has been missing, in spite of extensive efforts by structural biologists worldwide.

“We showed that, upon DNA binding, the conformation dynamics of the active domain site (HNH) of Cas9 triggers the formation of the active state, explaining how the HNH domain exerts a conformational control domain over DNA cleavage,” said Palermo. “Knowledge of the catalytic state greatly advances and helps in the effort of engineering CRISPR-Cas9 systems with improved specificity.”



## New Drug Candidate May Reduce Deficits in Parkinson's Disease

An international team led by UC San Diego researchers employed a novel computational approach to design and create a new compound that in laboratory studies reduced deficits and neurodegenerative symptoms that underlie Parkinson's disease.

In a study published in the September 27, 2016 Advance Access issue of *Brain*, the researchers describe how their compound, called NPT100-18A, prevents the binding and accumulation of alpha-synuclein or  $\alpha$ -syn in neuronal membranes, now considered a hallmark of Parkinson's disease and a related disorder called dementia with Lewy bodies.

"We've demonstrated a novel computational approach to design potential therapies for Parkinson's disease and related disorders," said the study's co-first author Igor Tsigelny, a research scientist with SDSC, the UC San Diego Moores Cancer Center, and the university's Department of Neurosciences.

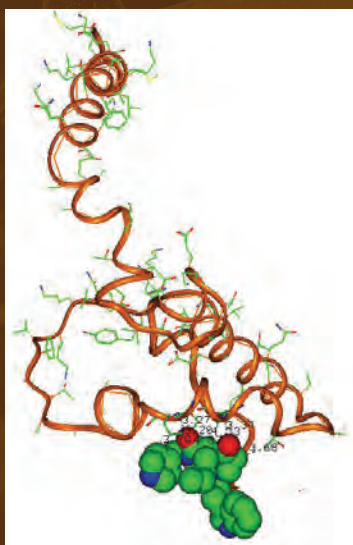
Parkinson's disease, which affects more than 10 million people worldwide, is characterized by impairment or deterioration of neurons in the area of the brain known as the *substantia nigra*. The disease typically occurs in people over the age of 60, with symptoms of shaking, rigidity, difficulty in walking, generally developing slowly over time and sometimes followed later by impairment in behavior and thought processes.

Since most symptoms of Parkinson's disease are triggered by a lack of dopamine in the brain, many medications are aimed at either temporarily replenishing dopamine or mimicking the action of this brain chemical. Unfortunately, current drugs have only a limited impact on long-term neurological deficits and mortality. For this reason, scientists have begun to focus their efforts on  $\alpha$ -syn's role in the disease, based largely on computer modeling showing how mutant forms of this protein penetrate and coil in cell membranes, and then aggregate in a matter of nanoseconds into dangerous ring structures that open pores to toxic ions that ultimately destroy neurons.

Enter several supercomputers – including SDSC's *Trestles*, *Gordon*, and the *Triton Shared Computing Cluster*, in addition to *Blue Gene* at the Argonne National Laboratory – that performed molecular dynamic simulations of *in silico* structures that would displace  $\alpha$ -syn from cell membranes.

Based on these simulations, other members of the research team, including Wolfgang Wrasidlo, executive director of medicinal chemistry at Neuropore Therapeutics in San Diego, synthesized a library of 34 potential compounds that targeted the "hot spot" where pairs of  $\alpha$ -syn proteins bind, merge, and aggregate in the cell membrane, an early step in the formation of toxic rings and ultimate death of a neuron. Of these drug candidates, the researchers identified one compound – NPT100-18A – as the most promising.

Funding for the research came from the National Institutes of Health, The Michael J. Fox Foundation for Parkinson's Research, and Neuropore Therapies.



IFT\_ASyn I8A5ns3: Interaction of the compound (colored by atoms: green-carbon, red-oxygen, blue-nitrogen) with alpha-synuclein (brown ribbon with color by atoms side chains).

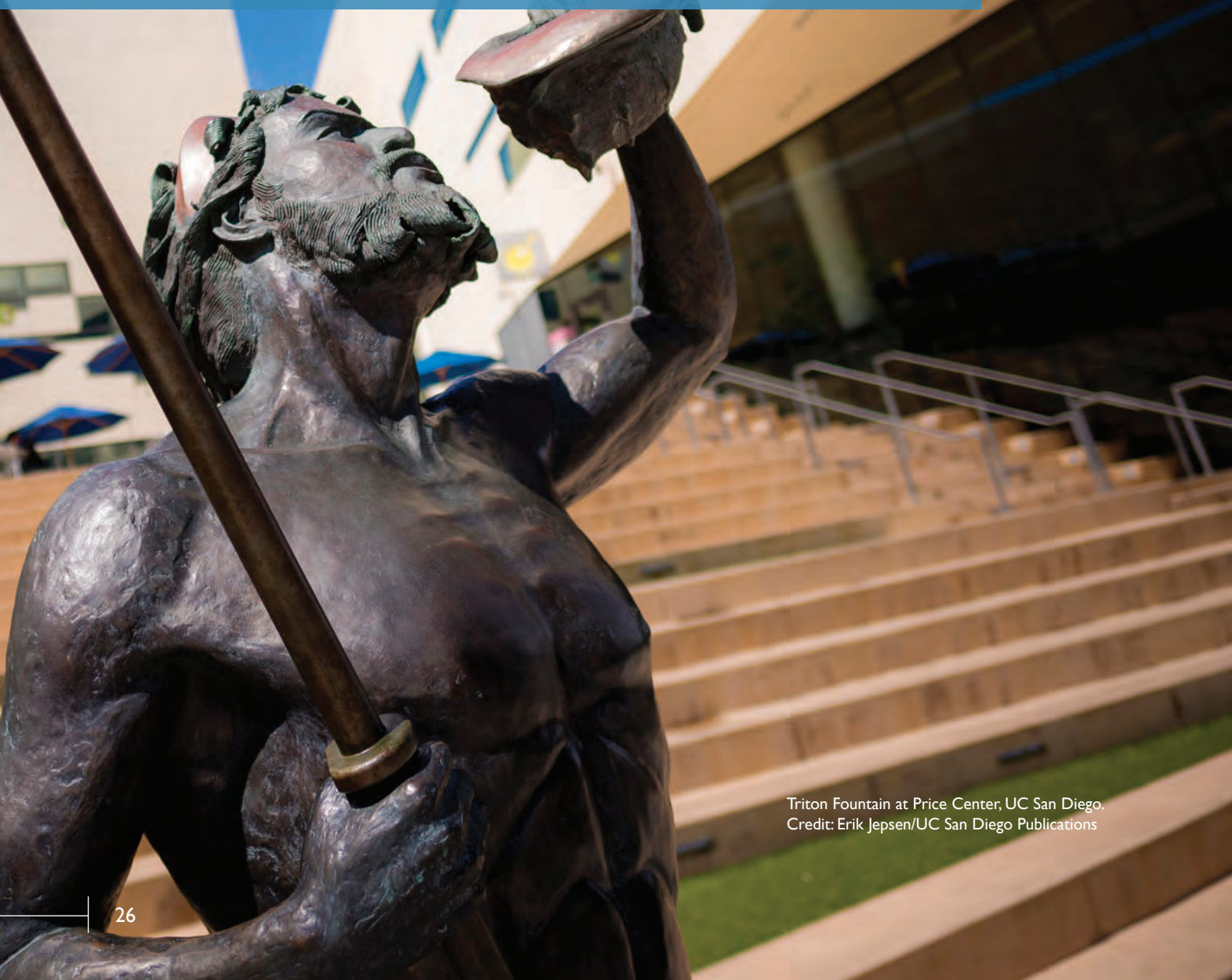
Credit: Igor Tsigelny, UC San Diego



IMPACT  
AND  
INFLUENCE

# LOCAL IMPACT & INFLUENCE

PROVIDING LEADING CYBERINFRASTRUCTURE  
AND EXPERTISE FOR UC SAN DIEGO'S  
GRAND RESEARCH CHALLENGES



Triton Fountain at Price Center, UC San Diego.  
Credit: Erik Jepsen/UC San Diego Publications

SDSC's dedication to harnessing meaning and value from a tidal wave of "big data" now generated by academic centers, commercial laboratories, government laboratories and observational tools is a major goal for UC San Diego researchers, as well as those in local industry and government. UC San Diego's strategic plan highlights four "grand research" themes of historical excellence at the university that can benefit from end-to-end "big data" cyberinfrastructure and educational know-how at SDSC. They include:

- Understanding and Protecting the Planet
- Enriching Human Life and Society
- Exploring the Basis of Human Knowledge, Learning, and Creativity
- Understanding Cultures and Addressing Disparities in Society

SDSC's development and operation of high-performance computing (HPC) resources at the national level provides substantial and tangible benefits to UC San Diego researchers, as well as the San Diego's burgeoning research communities. Specifically, SDSC's capabilities from its national standing as a center for an advanced research cyberinfrastructure support UC San Diego research areas the following areas:

### **'BIG DATA' SCIENCE**

Data from scientific research is being generated at a breathtaking pace, accelerated by the convergence of two new eras in the way research is now conducted: computational science, and data-intensive science and engineering – otherwise known as "Big Data" Science. SDSC is a pioneer in this field, and data-intensive science is at the core of the Center's new strategic plan.

### **SOFTWARE AND SUPPORT**

Related to big data, SDSC has developed original software for the national community that is also used by UC San Diego researchers. This includes Rocks cluster software; Kepler workflow management; Science Gateway infrastructure; virtual cluster software for integration with the Open Science Grid; the CIPRES, Neuroscience, and OpenTopography gateways; and the Protein Data Bank (read more about PDB developments on page 42).

### **BRAIN RESEARCH**

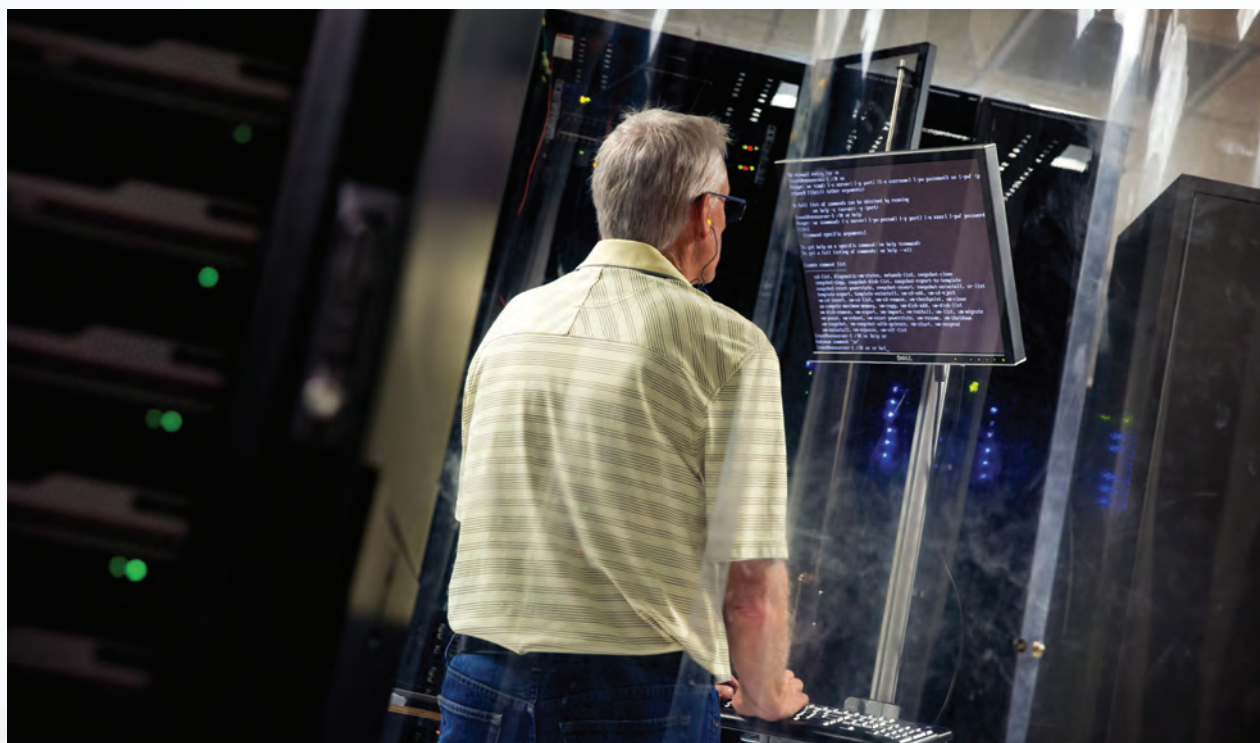
Much of this work now requires data-intensive computation using the most powerful supercomputers along with the ability to analyze huge volumes of information from this research. SDSC has the high-performance computing and big data resources, along with the technical expertise, to support this grand challenge. Read more in the Science Highlights section on page 19.

### **SMART SAN DIEGO**

Remote sensing networks, combined with the ability to assimilate vast quantities of information, not only will form the basis to forecast devastating earthquakes, but also will save lives and property from the ravages of wildfires. In recent years, SDSC has engaged in collaborative efforts with academic scientists at UC San Diego and elsewhere to develop computer codes and other systems that promise to dramatically reduce time and effort needed to simulate and respond to hazards.

### **EDUCATION, OUTREACH, AND TRAINING**

Includes numerous programs, conferences, and workshops such as SDSC's Summer Institute, International Conference on Computational Science, IEEE Women in Data Science Workshop, and others.



## CAMPUS COMPUTING CLUSTER GETS BIOINFORMATICS UPGRADES

In the modern research university, high-performance computing (HPC) is essential to advancing science in almost every discipline, from areas such as physics and engineering that are traditional users of HPC to disciplines such as economics and political science that are riding the “big data wave” and emerging as new users. SDSC’s contribution to research computing at UC San Diego has been the *Triton Shared Computing Cluster (TSCC)*, a “condo computing” program established in 2013. Condo computing is a shared ownership model in which researchers use funds from grants or other sources to purchase and contribute compute “nodes” to the system. The result is a researcher-owned, shared-campus computing resource of medium- to large-proportions, and much larger than could typically be afforded by the average researcher for dedicated use.

In early 2017, SDSC was awarded a National Science Foundation (NSF) grant to upgrade *TSCC* to deliver targeted capabilities for bioinformatics analyses. The grant, valued at almost half-a-million dollars and slated to run through January 2018, is part of the NSF’s Campus Cyberinfrastructure (CC\*) program, which invests in coordinated campus-level cyberinfrastructure (CI) components of data, networking, computing infrastructure, capabilities, and integrated services.

A key objective of the project is to leverage new technologies to provide accelerated computing capacity so that researchers can conduct thousands of additional whole-genome analyses per year, while also having the ability to conduct quick turnaround, single-genome analyses. The latter capability could be particularly useful for precision medicine and emerging clinical applications of genomics. Full details of the grant are in the Gateways to Discovery section on page 11.



To learn more about *TSCC* use the QR code or visit <https://goo.gl/wwaqL5>



HPWREN technicians install a microwave dish antenna on a remote mountaintop tower. Courtesy of HPWREN

## HPWREN: A WIRELESS EDUCATION AND SAFETY NETWORK FOR GREATER SAN DIEGO

As a collaborative project between SDSC and the Scripps Institution of Oceanography (SIO) at UC San Diego, the High Performance Wireless Research Network, or HPWREN was officially launched in early 2000 with NSF funding. HPWREN is an internet-connected “cyberinfrastructure” for research, education, and public safety that connects often hard-to-reach areas in remote environments via a system of cameras and weather stations to report local weather and environmental conditions, from severe rainstorms to wildfires and earthquakes. HPWREN also supports the Area Situational Awareness for Public Safety Network (ASAPnet), an extension of the HPWREN infrastructure for the benefit of public safety communities, especially firefighters in San Diego County. ASAPnet consists of a wireless internet data communications overlay supporting rural fire stations and other firefighter assets, in addition to environment-observing cameras and other sensors.

In 2016, HPWREN incorporated the AlertSoCal system, a network of mountaintop cameras operated by researchers at SIO, expanding Southern California’s state-of-the-art earthquake and weather monitoring systems to better detect fires in real time before they spread. New AlertSoCal 4K high-definition cameras augment the existing HPWREN cameras. AlertSoCal provides firefighters and the public with a virtual fire lookout tower equipped with real-time and on-demand time-lapse imagery up to 12 hours in the past to spot the first signs of fire ignition. The unprecedented view in these remote regions and within the wildland-urban interface can aid fire crews with critical information on fire evolution in its early stages to support safer operations and more timely evacuations of residents. Read more about HPWREN in the State Impact & Influence section starting on page 34.



To learn more about HPWREN use the QR code or visit <http://hpwren.ucsd.edu>



Ange Mason is SDSC's education program manager, helping to create innovative programs such as the Center's Research Experience for High School (REHS) students.



To learn more about REHS use a QR code reader or visit <https://goo.gl/mN5EWM>

# EDUCATING AND EMPOWERING THE NEXT GENERATION

Through its award-winning TeacherTech program established in 2001, SDSC has trained more than 2,500 teachers in the San Diego region in science and technology, helping many underserved students span the "digital divide" to the Information Age. SDSC also provides programs to train and educate local middle and high school students in computer science and technology, while fostering opportunities to bridge the gender gap for women in science, technology, education, and math (STEM). Called StudentTech, that program has engaged more than 3,000 students since its introduction in 2006.

## RESEARCH EXPERIENCE FOR HIGH SCHOOL STUDENTS

The Research Experience for High School Students (REHS) program, a part of SDSC's student outreach program, was developed to help increase awareness of computational science and related fields of research to students in the San Diego region. Students gain exposure to career options, hands-on computational experience, work-readiness skills, and mentoring by computational research scientists. Through the eight-week volunteer program conducted during the summer, students are paired with SDSC mentors to help them gain experience in an array of computational research areas. They learn how to formulate and test hypotheses, conduct computational experiments and draw conclusions from those experiments, and effectively communicate the science and societal value of their projects to a wide range of audiences. At the end of the program, students have a poster session to highlight their research and future career goals. More than 400 students have participated in REHS since it was started in 2010, and attendance levels have more than doubled in the last three years alone.



SDSC researchers Valentina Kouznetsova and Igor Tsigelny.

## Beyond REHS: Top Universities Open their Doors

Numerous students participating in SDSC's Research Experience for High School Students (REHS) annual summer program have not only continued their research projects but gone on to be accepted at top universities across the country.

Some notable examples come from SDSC researchers Igor Tsigelny and Valentina Kouznetsova, who specialize in computational drug design, personalized cancer medicine, gene networks analysis, and molecular modeling/molecular dynamics. Each year they mentor students in a variety of projects, and for 2017 added challenging artificial intelligence projects such as AI classification of beta-transcriptase inhibitors for finding new drugs to treat Alzheimer's disease.

Two-time REHS student Nathan Lian enrolled in Columbia University, while Alexandre Ettouati and Benjamin Li were accepted at UC Berkeley. In 2015, Tsigelny and Kouznetsova mentored nine students, and all of them were accepted at major universities including MIT, Caltech, Columbia, Northwestern, and Brown University. Those students include Eric Cunningham, Stephanie Hu, James Huang, Jane Huang, Nicholas Li, Andy Wang, Jonathan Wang, and Michelle Zhao.

In 2016, Tsigelny and Kouznetsova mentored 14 high school students in the REHS program. Each student had his or her own project. Three of them participated in the 2nd International Biomarker Conference in early 2017, while another participated in the 3rd Annual Public Health Research Day Symposium of UC San Diego last April.





## CALLING HIGH SCHOOL STUDENTS FOR UC SAN DIEGO'S MENTOR ASSISTANCE PROGRAM

San Diego-area high school students interested in pursuing a career in scientifically-based research are invited to apply to UC San Diego's Mentor Assistance Program (MAP), a campus-wide initiative designed to engage students in a mentoring relationship with an expert from a vast array of disciplines. The mentoring period runs from September 2017 through May 2018.



Launched about two years ago by SDSC and UC San Diego School of Medicine, MAP's mission is to provide a pathway for student researchers to gain access to UC San Diego faculty, post-doctoral fellows, Ph.D. candidates, and staff to mentor them in their own field of interest. Mentors are recruited from across campus from fields that include athletics, biology, chemistry, aerospace engineering, network architectures, pharmaceutical sciences, physics, social studies, and more.

Poster sessions for SDSC's REHS Program (top) and MAP programs let students share their work with educators and their families.

"MAP is an opportunity for students to take the first step into a potential career path, while simultaneously building an early foundation for success in their academic career," said SDSC Education Manager and MAP co-founder Ange Mason. "These mentoring relationships are intended to support collegiality, effective communication, self-evaluation, and cultural competence, all of which enhance a stimulating and supportive university environment."



To learn more about MAP use the QR code or visit <https://goo.gl/s4wMVB>

Research Experience for High School Students (REHS) Class of 2017.



# STATE IMPACT & INFLUENCE

ALIGNING WITH  
PRINCIPLES AND PARTNERSHIPS



# UC@SDSC

## DATA-ENABLED SCIENCE BASED ON COLLABORATION, INNOVATION, & EDUCATION

In 2014, SDSC launched an initiative called UC@SDSC – an engagement strategy that highlights collaboration, innovation, and education while promoting the Center’s resources and technical expertise as a valuable asset across the entire UC system. An External Advisory Board consisting of Vice Chancellors, Deans, and Distinguished Professors from all 10 UC campuses and leadership from three national laboratories was formed to provide guidance and recommendations, resulting in numerous new collaborations. Highlights include:

### HPC@UC

Launched in mid-2016, HPC@UC provides UC researchers with access to SDSC’s high-performance computing resources including *Comet* as well as the expertise required to make efficient use of them. The collaboration is offered in partnership with the UC Vice Chancellors of Research as well as campus CIOs. HPC@UC is intended to broaden the base of UC researchers who use advanced computing while seeding promising computational research. As of June 2017, SDSC awarded nearly 6 million core-hours to some 23 projects from researchers at UC Santa Barbara, UC Irvine, UC Santa Cruz, UC Los Angeles, UC San Diego, UC Davis, UC Berkeley, and UC San Francisco.



To read more about HPC@UC use a QR code reader or visit [www.sdsc.edu/collaborate/hpc\\_at\\_uc.html](http://www.sdsc.edu/collaborate/hpc_at_uc.html)

### HPC WORKSHOPS

SDSC staff conducted numerous High-Performance Computing/Big Data workshops at UC Davis, UC Irvine, UCLA, and UC Santa Barbara, attracting hundreds of attendees that included graduate students, postdocs, and faculty from various departments, along with technical staff. SDSC’s instructors, who hold Ph.Ds. in physics, astrophysics, aerospace engineering, computer science, and cognitive science, visited the campuses to teach these one-day tutorials on the latest data-intensive technologies and how they combine with HPC and scientific computing. These workshops, which promote good interaction among researchers from those UC campuses, have been conducted since the start of the UC@SDSC program in 2014, and there is strong interest for them to continue.

SDSC also recently completed a collaboration through XSEDE’s Extended Collaborative Support Service, with Maria-Grazia Ascenzi, associate research scientist at UCLA/Orthopaedic Hospital’s Department of Orthopaedic Surgery. The collaboration, which ran from July 2016 to March 2017, focused on automation of a hierarchical finite element modeling pipeline of a human femur bone using Abaqus and custom Python

interpolation functions. Thanks to the collaboration, the pipeline execution time was sped up by a factor of three and some software bugs were identified and fixed.



To read more about HPC Workshops use a QR code reader or visit <https://goo.gl/TrLqK9>

### PACIFIC RESEARCH PLATFORM

To meet the needs of researchers in California and beyond, the National Science Foundation awarded a five-year grant to fund the Pacific Research Platform (PRP). The PRP’s data-sharing architecture, with end-to-end 10-100 gigabits per second (Gb/s) connections, will enable region-wide virtual co-location of data with computing resources and enhanced security options.

Led by Calit2 Director Larry Smarr, Calit2 Researcher Tom DeFanti; SDSC’s Frank Würthwein and Phil Papadopoulos; John Graham (UC San Diego); Camille Crittenden (UC Berkeley); John Hess (CENIC), Thomas Hutton (SDSC) and Eli Dart (ESnet); the PRP supports a broad range of data-intensive research projects that will have wide-reaching impacts on science and technology worldwide. Cancer genomics, human and microbiome ‘omics’ integration, biomolecular structure modeling, galaxy formation and evolution, telescope surveys, particle physics data analysis, simulations for earthquakes and natural disasters, climate modeling, virtual reality and ultra-resolution video development are just a few of the projects that are benefiting from the use of the PRP. The PRP will be extensible across other data-rich domains as well as other national and international networks potentially leading to a national and eventually global data-intensive research cyber-infrastructure.



To read the full press release about the PRP use a QR code reader or visit <https://goo.gl/asXPx7>



(Above) View of the Whittier Fire in Santa Barbara County as seen from one of HPWREN's antenna-mounted observation cameras on top of Santa Ynez Peak. Courtesy of HPWREN/SDSC

## IMPACT OF HPWREN'S WIRELESS EDUCATION AND SAFETY NETWORK IS FELT STATEWIDE

Wildfire activity throughout California during 2017 reached new levels as the season extended into mid-December with devastating consequences. In Northern California, wildfires in October destroyed entire neighborhoods in parts of Sonoma and Napa counties, with more than 30 lives lost along with some 3,500 homes and businesses. Fires in Ventura and Santa Barbara counties scorched more than 130,000 acres, with more than 400 homes and structures destroyed by blazes whipped by Santa Ana winds. In early December, some 46 horses were killed at a thoroughbred training facility during the Lilac wildfire in northern San Diego County.

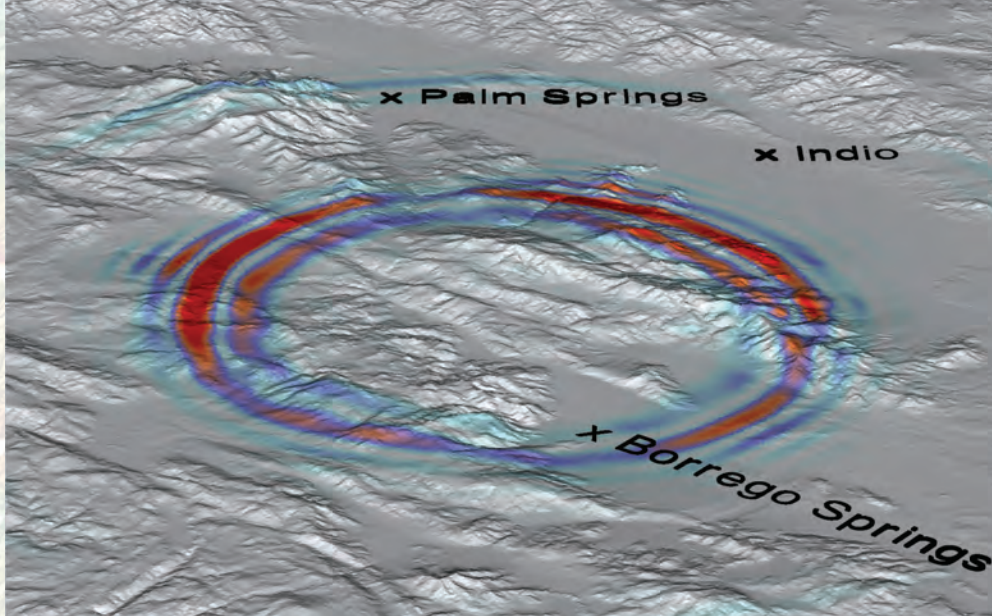


Research geophysicist Frank Vernon points to streaming imagery from the AlertSoCal network at Scripps Institution of Oceanography. Credit: Richard Klein

Firefighters in greater San Diego, as well as those in Santa Barbara, were aided by UC San Diego's High Performance Wireless Research Network (HPWREN), established in 2000 under a collaboration between SDSC and the Scripps Institution of Oceanography at UC San Diego. In late 2016, HPWREN installed a set of cameras on top of Santa Ynez Peak in Santa Barbara County in partnership with Jamison Steidl, a researcher at UC Santa Barbara's Earth Research Institute. The system of fixed cameras and a pan-tilt-zoom (PTZ) camera was configured to show the extensibility of HPWREN services across the county and into neighboring counties.

In July 2017, that network was put to a real-world test when the Whittier Fire in Santa Barbara County broke out, burning more than 18,000 acres. Incident commanders were able to remotely direct and focus the PTZ camera on areas of immediate interest as they sought to contain the wildfire.

The transition from viewing individual images to watching real-time streaming video allowed responders to better see the effects of wind, humidity, and temperature on the spread of the fire, according to HPWREN co-founder Frank Vernon. "Feedback was extremely positive and many of the new capabilities we provided ad-hoc during the emergency are now being designed into our systems, along with mechanisms to more easily manage them and respond to requests during future emergencies."



## New Software and Latest Intel Processors Help to Better Predict Earthquake Paths

SDSC researchers recently developed a new seismic software package with Intel Corporation that has enabled the fastest seismic simulation to-date, as the two organizations collaborate on ways to better predict ground motions to save lives and minimize property damage.

The new simulations, which mimic possible large-scale seismic activity in the southern California region, were done using a new software system called EDGE, for Extreme-Scale Discontinuous Galerkin Environment. “Obtaining higher frequencies is a key to predicting ground motions relevant to common dwellings as part of conducting earthquake research,” said Yifeng Cui, director of the High Performance Geocomputing Laboratory (HPGeoC) at SDSC.

“Our research also showed a substantial gain in efficiency in using the new software,” said Alex Breuer, a postdoctoral researcher and lead author of a paper on the study that was presented in June 2017 at the International Super Computing (ISC) High Performance conference in Frankfurt, Germany. “Researchers will be able to run about two to almost five times the number of simulations using EDGE, saving time and reducing cost.”

A second HPGeoC paper presented at the ISC High Performance conference covered a new study of the AWP-ODC software used for years by co-collaborators at the Southern California Earthquake Center (SCEC). The software was optimized to run in large-scale for the first time on the latest generation of Intel data center processors known as Intel Xeon Phi x200.

Such detailed computer simulations allow researchers to study earthquake mechanisms in a virtual laboratory. “These two studies open the door for the next-generation of seismic simulations using the latest and most sophisticated software,” said Cui. “Going forward, we will use the new codes widely for some of the most challenging tasks at SCEC.”

Both research projects are part of a collaboration announced in early 2016 under which Intel opened a computing center at SDSC to focus on seismic research, including the ongoing development of computer-based simulations that can be used to better inform and assist disaster recovery and relief efforts. In addition to UC San Diego, the Intel PCC at SDSC includes researchers from the University of Southern California (USC), San Diego State University (SDSU), and the University of California Riverside (UCR).



For more information on SDSC's new seismic simulations, use a QR code reader or visit <https://goo.gl/RyB3fO>



# NATIONAL IMPACT & INFLUENCE

SDSC'S NATIONAL MISSION IN ADVANCED  
CYBERINFRASTRUCTURE

As one of the country's first four supercomputer centers opened in 1985 by the National Science Foundation (NSF), SDSC has a long history of programs and partnerships that have benefited science and society across a wide variety of science domains.

SDSC's mission has expanded with time to encompass more than just advanced computation, which has served as a foundation to include new and innovative applications and expertise related to the ever-increasing amount of digitally-based science data generated by researchers.

"Today, data-enabled science is inextricably linked to, and based upon, computationally-based research and expertise," said SDSC Director Michael Norman. "SDSC's standing as a well-regarded national supercomputing center, along with the Center's commitment to fulfilling its national advanced cyberinfrastructure mission, provide significant benefits to UC San Diego, UC researchers, and a wide range of industry users. Such partnerships are essential to advancing scientific discovery aimed at solving the grand research challenges of our times."

SDSC has key national partnerships with the following programs:



## EXTREME SCIENCE AND ENGINEERING DISCOVERY ENVIRONMENT (XSEDE)

The National Science Foundation's XSEDE program provides academic researchers with the most advanced collection of integrated advanced digital resources and services in the world. As the only supercomputer center participant on the West Coast, SDSC provides advanced user support and expertise for XSEDE researchers across a variety of applications. SDSC's *Comet* supercomputer is accessible via the XSEDE allocation process to U.S. researchers as well as those affiliated with U.S.-based research institutions. *Comet* is among the most widely used systems in XSEDE's resource portfolio.



Frank Würthwein is the Executive Director of the Open Science Grid and SDSC's lead for distributed high-throughput computing.

## THE OPEN SCIENCE GRID CONSORTIUM

The Open Science Grid (OSG) is a multi-disciplinary research partnership specializing in high-throughput computational services funded by the U.S. Department of Energy and the NSF. Through a partnership with XSEDE, OSG scientists have access to resources such as *Comet* to further their research. The integration of *Comet* into the OSG provisioning system was led by a team including Frank Würthwein, an expert in experimental particle physics and advanced computation and SDSC's lead for distributed high-throughput computing. Würthwein served as OSG's Executive Director during FY2016/17. OSG operates services that allow for transparent computation across more than 150 computing clusters worldwide, including National Grid Initiatives in Europe, Asia, and the Americas.

## NSF WEST BIG DATA INNOVATION HUB (WBDIH)

The National Science Foundation supports four regional Big Data Innovation Hubs throughout the U.S. The Western region is comprised of 13 states with Montana, Colorado, and New Mexico marking the eastern boundary. The WBDIH is led from SDSC, UC Berkeley, and the University of Washington's eScience Institute. The Hub's purpose is to connect, educate, incubate, and facilitate multi-state, multi-sector partnerships in the area of big data innovation. Thematic spokes include Managing Natural Resources & Hazards, Metro Data Science, and Precision Medicine with supporting thematic rings that include Big Data Technology and Data-Enabled Scientific Discovery & Learning. Current spoke projects include MetroInsight, a 'smart' and connected city project led out of UC San Diego; CoMSES Net (Network for Computational Modeling in Social and Ecological Sciences), a spoke dedicated to promoting and enabling open and reproducible scientific computation led by the Arizona State University; Big Data and Criminal Justice, led from Boise State; and a planning grant led by the Institute for Systems Biology and SDSC to increase collaborations in proteogenomics applications of genetic variations. The WBDIH regularly hosts workshops including recent ones such as Data Storytelling, Data Hackathon Best Practices, and the National Transportation Data Challenge.



To learn more about WBDIH use the QR code or visit [westbigdatahub.org](http://westbigdatahub.org)

## SUPPORTING THE NATIONAL BRAIN INITIATIVE THROUGH THE NEUROSCIENCE GATEWAY

Charting brain functions in unprecedented detail could lead to new prevention strategies and therapies for disorders such as Alzheimer's disease, schizophrenia, autism, epilepsy, traumatic brain injury, and more. The BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies), launched by President Barack Obama in 2013, is intended to advance the tools and technologies needed to map and decipher brain activity, including advanced computational resources and expertise.



Christine Kirkpatrick is the executive director of the National Data Service and division director for IT Systems & Services at SDSC.

## NATIONAL DATA SERVICE (NDS)

SDSC has taken a leadership role in the burgeoning National Data Service through its appointment of SDSC's Christine Kirkpatrick as the organization's first executive director. The NDS is a U.S. consortium of research computing centers, governmental agencies, libraries, publishers and universities. NDS builds on the data archiving and sharing efforts already underway within scientific communities and links them together with a common set of services. NDS's unique value-add is not in new tools but in making it easier to use such tools together through dependency management and increased interoperability between cyberinfrastructure services. NDS is an emerging vision for how scientists and researchers across all disciplines can find, reuse, and publish data, while providing a platform and 'sandbox' for developers creating data services.



To learn more about NDS use the QR code or visit [www.nationaldataservice.org](http://www.nationaldataservice.org)



## DATA SCIENCE AND BIG DATA COURSES AVAILABLE ONLINE

SDSC and UC San Diego recently launched a four-part Data Science series via edX's MicroMasters® program. In partnership with Coursera, SDSC launched a series of MOOCs (massive open online courses) as part of a Big Data Specialization that has proven to be one of its top course series. Consisting of five courses and a final Capstone project, this specialization provides valuable insight into the tools and systems used by big data scientists and engineers. In the final Capstone project, learners apply their acquired skills to a real-world big data problem. To date, the courses have reached more than 700,000 students in every populated continent – from Uruguay to the Ivory Coast to Bangladesh. A subset of students pay for a certificate of completion.



To learn more about SDSC's Data Science and Big Data courses offered through the Coursera platform, use a QR code reader or visit [www.coursera.org/specializations/bigdata](http://www.coursera.org/specializations/bigdata)



## SDSC HIGH-PERFORMANCE COMPUTING (HPC) SUMMER INSTITUTE

SDSC's Summer Institute is an annual week-long training program offering introductory to intermediate topics on high-performance computing and data science. The theme for SDSC's 2016 HPC Summer Institute focused on the "long tail of science," or the idea that the large number of modest-sized, computationally-based research projects still represents, in aggregate, a tremendous amount of research and resulting scientific impact. The week-long event, which allowed attendees to perform hands-on exercises on SDSC's *Comet* supercomputer, included plenary sessions covering essential skills including data management, running jobs on SDSC resources, and various techniques for turning data into meaningful and usable knowledge. SDSC recently graduated its 2017 Summer Institute group representing 31 institutions from around the U.S. The program was expanded to cover new topics, such as Machine Learning at Scale, distributed programming in Python, cluster computing with Spark, and CUDA programming. SDSC has conducted the Summer Institute since the mid-1990s.



To learn more about the SDSC Summer Institute use a QR code reader or visit [www.sdsc.edu/events/summerinstitute/](http://www.sdsc.edu/events/summerinstitute/)

# NATIONAL IMPACT & INFLUENCE

## PROVIDING 'SCIENCE GATEWAYS' FOR RESEARCHERS



Nancy Wilkins-Diehr is an associate director of SDSC and co-PI of XSEDE (Extreme Science and Engineering Discovery Environment) as well as 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.

Surpassing the 10,000-user milestone in less than two years of *Comet's* operations (see details on page 8) was due in large part to researchers accessing the resource via science gateways, which provide scientists with access to many of the tools used in cutting-edge research – telescopes, seismic shake tables, supercomputers, sky surveys, undersea sensors, and more – connecting often diverse resources in easily accessible ways that save researchers and institutions time and money.

Science gateways make it possible to run the available applications on supercomputers such as *Comet* so results come quickly, even with large data sets. Moreover, browser access offered by gateways allows researchers to focus on their scientific problem without having to learn the details of how supercomputers work and how to access and organize the data needed. SDSC alone has delivered 77 percent of all gateway cycles since the start of the XSEDE project in 2011.

In mid-2016, a collaborative team led by SDSC Associate Director Nancy Wilkins-Diehr was awarded a five-year, \$15 million NSF grant to establish a Science Gateways Community Institute (SGCI) to accelerate the development and application of highly functional, sustainable science gateways that address the needs of researchers across the full spectrum of NSF directorates.

“It’s possible to support gateways across many disciplines because of the variety of hardware and support for complex, customized software environments on *Comet*,” said Wilkins-Diehr, who also is co-director of XSEDE’s Extended Collaborative Support Services. “This is a great benefit to researchers who value the ease of use of high-end resources via such gateways.”

In November 2016, SGCI held its first annual conference, Gateways 2016, which was hosted by SDSC. The conference attracted almost 120 attendees and about 40 submissions for tutorials, paper presentations, demos, and panels. In May 2017, as part of its education and outreach efforts, SGCI hosted three programs for undergraduate and graduate students, with programs taking place at various SGCI partner sites. Each program included a stipend, travel support, and housing for students and was quickly oversubscribed. This year’s programs included:

- A three-day Science and Engineering Applications Grid (SEAGrid) computational chemistry workshop, hosted by Jackson State University in Mississippi, that included an opportunity for students to move their research project onto the grid. A four-week program focused on gateway development for undergraduate students at Elizabeth City State University, in North Carolina. The workshop covered the core skills needed to be productive in design and maintenance of science gateways.
- An eight-week internship for students working on a gateway project. Participants were placed at one of the SGCI partner sites. The program was open to graduate students and undergraduates who have completed their junior year with majors in computer science or computer engineering related fields with strong programming and software engineering skills.

In all, some 32 science gateways are available via XSEDE’s resources, each one designed to address the computational needs of a particular community such as computational chemistry, phylogenetics, and the neurosciences.

# SCIENCE GATEWAYS PIONEERED BY SDSC RESEARCHERS



For more information on CIPRES, use a QR code reader or visit [www.phylo.org](http://www.phylo.org)

## CIPRES

One of the most popular science gateways across the entire XSEDE resource portfolio is the CIPRES science gateway, created as a portal under the NSF-funded CyberInfrastructure for Phylogenetic REsearch (CIPRES) project in late 2009. The gateway is used by scientists to explore evolutionary relationships by comparing DNA sequence information between species. To date, the CIPRES science gateway has supported more than 23,000 users conducting phylogenetic studies involving species in every branch of the “tree of life”. The gateway is used by researchers on six continents, and their results have appeared in more than 3,500 scientific publications since 2010, including *Cell*, *Nature*, and *PNAS*.

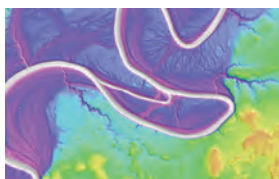
“The scheduling policy on *Comet* allows us to make big gains in efficiency because we can use anywhere between one and 24 cores on each node,” said Mark Miller, principal investigator of the CIPRES gateway and an SDSC biologist. Typically, about 200 CIPRES jobs are running simultaneously on *Comet*. “When you are running 200 small jobs 24/7, those savings really add up in a hurry.” Read more about CIPRES-enabled research on page 17 in the Biodiversity Science Highlight.



For more information on NSG, use a QR code reader or visit [www.nsgportal.org](http://www.nsgportal.org)

## NEUROSCIENCES GATEWAY (NSG)

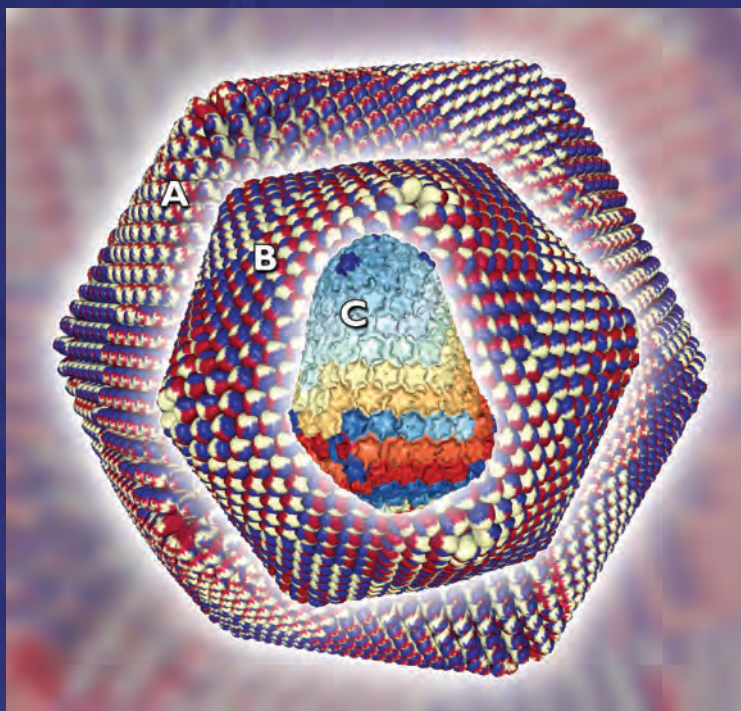
Computational modeling of cells and networks has become an essential part of neuroscience research, and investigators are using models to address problems of ever-increasing complexity. The Neuroscience Gateway (NSG) provides researchers with access to the popular computational neuroscience tools installed on various high-performance computing (HPC) resources funded by the NSF, along with a community “mailing list” for neuroscientists to collaborate and share ideas. In 2015, the NSF and the United Kingdom’s Biotechnology and Biological Sciences Research Council (BBSRC) awarded funding for a new Neuroscience Gateways project led by SDSC. That project, which will contribute to the national BRAIN initiative, is a collaboration between UC San Diego, with SDSC’s Amit Majumdar as the principal investigator (PI) and Subhashini Sivagnanam as co-PI; Yale University, with Ted Carnevale as PI; and University College London, with Angus Silver as PI. Read more about the NSG’s role in advancing neuroprosthesis research on Page 19.



For more information on OpenTopography use a QR code reader or visit [www.opentopography.org](http://www.opentopography.org)

## OPENTOPOGRAPHY

Initiated in 2009 with funding from the NSF, OpenTopography provides easy access to earth science-oriented, high-resolution topographical data and processing tools for a broad spectrum of research communities. A collaboration between UC San Diego, Arizona State University and UNAVCO, OpenTopography employs sophisticated cyberinfrastructure that includes large-scale data management, high-performance computing (HPC), and service-oriented architectures, providing researchers with efficient web-based access to large, high-resolution topographic datasets. Currently, OpenTopography data holdings comprise 253 lidar point cloud datasets with over one trillion lidar returns and 145 high-resolution raster datasets covering 145,992 km<sup>2</sup>, and four global datasets including the highly popular Satellite Radar Topography Mission (SRTM) global 30m dataset. The OpenTopography gateway has been averaging close to 400 new user registrations per month since the start of 2016, with more than 18,768 registered users and numerous others accessing data and running jobs as guests. In 2016 alone, 16,459 unique users ran over 53,192 jobs via the portal, and an additional 171,496 jobs were invoked through its available Application Program Interfaces.



Cross-sections of three virus capsids overlaid for size comparisons: A) Faustovirus, PDB ID 5J7V, ~40 million atoms, B) PBCV-1 virus, PDB ID 1M4X, ~16 million atoms, C) HIV-1 virus, PDB ID 3J3Q, ~ 2.4 million atoms, displayed in a web browser using NGL viewer and MMTF technology. Image credit: Ben Tolo, SDSC



Peter Rose heads SDSC's Structural Bioinformatics Laboratory and leads bioinformatics and biomedical applications for the Center's Data Science Office.

## Structural Biology Enters the Big Data Era

SDSC's Structural Bioinformatics Laboratory directed by Peter Rose collaborated with the RCSB Protein Databank to develop a highly efficient representation of 3D biomolecular structures such as proteins, DNA, and RNA. The visualization and analysis of these 3D structures are essential to elucidate the atomic-level detail of biological processes and explore the mechanism of action of drug molecules. However, rapid growth and increasing molecular complexity in the PDB has created challenges for the interactive exploration of these structures. Funded by the National Institutes of Health's Big Data to Knowledge (BD2K) initiative, the lab developed the MacroMolecular Transmission Format (MMTF) for the rapid network transfer and processing of 3D structures. The results of these developments were recently published in *PLOS Computational Biology*. Using this new technology, even the largest structures with millions of atoms can be displayed in a web-browser on a desktop or laptop computer as well as mobile devices. The same technology is being developed to increase the speed of PDB data mining. "By combining our efficient 3D macromolecular representation with parallel processing using Apache Spark, we can mine the PDB of more than 130,000 structures in minutes, rather than hours or days using traditional approaches," said Rose.



For more information about PDB MMTF, use a QR code reader or visit <http://mmtf.rcsb.org>



FOCUSED  
SOLUTIONS and  
APPLICATIONS



# FOCUSED SOLUTIONS AND APPLICATIONS

As one of the country's first four supercomputer centers opened in 1985 by the National Science Foundation (NSF), SDSC has a long history of programs and partnerships that have benefited science and society across a wide variety of science domains.

SDSC's wide range of expertise in advanced computation and data-enabled science has yielded collaborations at the local, state, national, and even international levels. Many of these partnerships bring together researchers across academia, industry, and government to advance scientific discovery ranging from deepening our understanding of how the human brain works and developing new drugs to fight debilitating diseases to creating detailed simulations to help relief services better cope with natural and man-made disasters around the world.

SDSC's resources and expertise are organized to support UC San Diego's strategic plan and research themes which include: understanding and protecting the planet; enriching human life and society; exploring the basis of human knowledge, learning, and creativity; and understanding cultures and improving society: addressing disparities, social justice, access, equity and inclusion. SDSC also continues to align itself with UC principles that include increased integration of the Center's activities with researchers and faculty across multiple UC campuses; enhancing the efficiency and reputation of UC; and providing innovation and discovery to help improve the health and welfare of the citizens of California.

SDSC's new strategic plan, under development for the better part of a year, sets forth three key focus areas that allow us to support these research themes:

- ADVANCED, VERSATILE COMPUTING SYSTEMS
- DATA-DRIVEN SCIENCE PLATFORMS & APPLICATIONS
- LIFE SCIENCE COMPUTING & APPLICATIONS

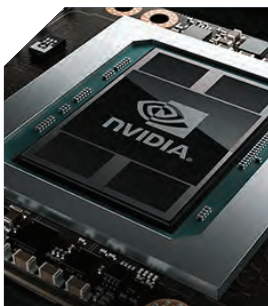
These areas embody broad themes that will chart SDSC's course over the next several years while connecting to national, state, and local priorities of critical importance. Some highlights that support each priority appear on the pages that follow in this section.

# FOCUSED SOLUTIONS FOR ADVANCED, VERSATILE COMPUTING SYSTEMS



## EXPLORING EMERGING TECHNOLOGIES

Through SDSC's Advanced Technology Lab (ATL), researchers are evaluating new computing hardware and software technologies to better understand how they can be used in future computing systems for the national research community, while ensuring that SDSC remains competitive in securing these systems. Working closely with the private sector and other organizations, ATL researchers gain access to the newest technologies, in many cases before they are generally available or while still under development. One recent ATL project includes looking into data movement at a lower level within processor architectures, as this is becoming a determining factor in processor architectures from performance and energy aspects. Another project involves performance characterization of biosciences applications on multi-core processors for widely used bioinformatics and cryo-electron microscopy data analysis software. Other areas of research include hierarchical storage systems (e.g. burst buffer technologies), non-volatile memory, and emerging architectures. SDSC's ATL has been in operation just over a year and has received funding and in-kind contributions from various industries and organizations.



## DOUBLING DOWN ON COMET'S GPU POWER

In early 2017, SDSC was granted a supplemental award from the NSF to double the number of graphic processing units, or GPUs, on its petascale-level *Comet* supercomputer in direct response to growing demand for GPU computing across a wide range of research domains. Under the supplemental award, valued at just over \$900,000, SDSC expanded *Comet* with the addition of 36 GPU nodes, each with four NVIDIA P100s, for a total of 144 GPUs – and doubling the total number of GPUs to 288. The expansion makes *Comet* the largest provider of GPU resources available to the NSF-funded Extreme Science and Engineering Discovery Environment (XSEDE), a national partnership of institutions that provides academic researchers with the most advanced collection of digital resources and services in the world. Read more about *Comet* in SDSC's Gateways to Discovery section on page 8.



## GORDON LIVES!

SDSC's *Gordon* supercomputer, which has served as a key NSF/XSEDE resource for five years, continues to advance scientific discovery under a new partnership with the Simons Foundation's Flatiron Institute in New York. Under a multi-year agreement, the majority of the data-intensive supercomputer is now being used by the Institute for ongoing research in astrophysics, biology, condensed matter physics, materials science, and other domains. SDSC has retained a smaller portion of *Gordon* for use by other organizations including UC San Diego's Center for Astrophysics & Space Sciences (CASS), as well as SDSC's OpenTopography project and various projects within the Center for Applied Internet Data Analysis (CAIDA), which is based at SDSC. Read more about this agreement in the Gateways to Discovery section on page 10.



## ENHANCING BIOINFORMATICS ANALYSES

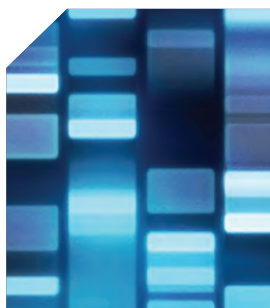
In late 2016, SDSC launched an initiative focused on improving the overall performance of bioinformatics applications and associated analysis pipelines on current and future advanced computing systems. Then in early 2017, SDSC was awarded an NSF grant to upgrade UC San Diego's *Triton Shared Computing Cluster (TSCC)* to enhance capabilities for bioinformatics analyses. That grant, slated to run through January 2018, is part of the NSF's Campus Cyberinfrastructure (CC\*) program, which invests in coordinated campus-level cyberinfrastructure (CI) components of data, networking, computing infrastructure, capabilities, and integrated services. Read more about these awards under Life Sciences Applications: *TSCC* on the next page.



# FOCUSED SOLUTIONS FOR LIFE SCIENCE COMPUTING & APPLICATIONS

This strategic priority addresses emergent computing and ‘big data’ challenges in life sciences. Elements include:

- Development of an integrated environment for configuring, testing, and validating new informatics platforms, software stacks, workflows, and other technologies in support of commercial, non-profit, and academic-based life science and biomedical research;
- Assessment of current capabilities and gaps for integration of protected health data systems with computational platforms;
- Utilization of advanced sensing technologies for biomedical applications; and
- Benchmarking and performance optimization of bioinformatics applications.



## ADVANCES IN BIOINFORMATICS APPLICATIONS & ANALYSES

Spurred by the increasing reliance of life sciences researchers in the academic and private sectors on computational methods and data-enabled science, SDSC inaugurated a new life sciences computing initiative in 2016 focused on improving the performance of bioinformatics applications and related analyses on advanced computing systems. The initial work, co-sponsored and supported by Dell and Intel, centers on benchmarking selected genomic and Cryo-electron Microscopy (Cryo-EM) analysis pipelines and developing recommendations for technical architectures to service those pipelines. Such recommendations include integrated computing and storage platforms as well as networking fabrics.

Dramatic improvements in scientific instruments and techniques, such as Next Generation Sequencing (NGS) and Cryo-EM, are enabling the rapid accumulation of vast amounts of data including DNA/RNA molecular sequences and high-resolution imaging of biological structures from animal and plant organisms. “Our experience with both on-campus researchers and biotech companies is that refined bioinformatics techniques and new technologies can be leveraged to improve the breadth and throughput of analyses,” said Wayne Pfeiffer, an SDSC Distinguished Scientist and the Center’s bioinformatics lead.

SDSC’s initiative focuses on developing and applying rigorous approaches to assessing and characterizing computational methods and pipelines. It will then specify the architectures, platforms, and technologies to optimize performance and throughput, among other dimensions



## TSCC: A SHARED CAMPUS CLUSTER

In early 2017, SDSC was awarded an NSF grant to upgrade UC San Diego’s *Triton Shared Computing Cluster (TSCC)* to deliver targeted capabilities for bioinformatics analyses. That grant, valued at almost half-a-million dollars and slated to run through January 2018, is part of the NSF’s Campus Cyberinfrastructure (CC\*) program, which invests in coordinated campus-level cyberinfrastructure (CI) components of data, networking, computing infrastructure, capabilities, and integrated services. A key objective of the project is to leverage new technologies to provide accelerated computing capacity so that researchers can conduct high throughput analyses of many whole genomes, while also having the ability to conduct quick turnaround, single-genome analyses. The latter capability could be particularly useful for precision medicine and emerging clinical applications of genomics. Read more about this award in Gateways to Discovery section on page 11.

# FOCUSED SOLUTIONS FOR DATA SCIENCE PLATFORMS & APPLICATIONS

SDSC has been a leader in big data applications and systems that relate to scalable management and processing of diverse streaming and batch datasets well before the term ‘big data’ was coined. Several recent projects underscore SDSC’s leadership in this strategic priority area:

- WIFIRE ([wifire.ucsd.edu](http://wifire.ucsd.edu)) as a big data application as well as a streaming and processing platform development with societal impact in incident management, real-time prediction, and decision-making support for wildfires;
- WBDIH ([westbigdatahub.org](http://westbigdatahub.org)) as a big data hub to facilitate multisector collaboration among the 13 western states to address research challenges in areas such as precision medicine, natural resource and hazard management, and metro data science;
- LHC ([home.cern/topics/large-hadron-collider](http://home.cern/topics/large-hadron-collider)) as a big data platform with an ability to handle massive amounts of data streaming from particle accelerators at unprecedented velocity; and
- AWESOME (Analytical WorkBench for Social Media Analytics) as a big data platform to continuously ingest multiple sources of real-time social media data and scalable analysis of such data for applications in social science, digital epidemiology, and internet behavior analysis.



## DATA SCIENCE HUB

More than two years in the making, SDSC recently established a conceptual framework for a “Data Science Hub” (DSH) where experts from SDSC and other areas of UC San Diego can apply their experience to create collaborative, multi-disciplinary data science teams to help provide solutions to regional, national, and global challenges such as ‘smart’ cities, precision medicine, advanced manufacturing, and data center automation.

Key goals of the Data Science Hub framework include:

- Serving as a community hub for collective, lasting innovation in data science research and expertise;
- Leading innovative solutions and applications for data-driven research, analytics, and development;
- Creating top-of-the-line computing and data platforms for the campus data science community;
- Educating and establishing a modern data science workforce that can help drive innovation in public and private organizations; and
- Connecting data science research initiatives with entrepreneurial ventures and potential industry commercialization.



For more information on the Data Science Hub at SDSC use a QR code reader or visit <http://datascience.sdsc.edu>

With a \$75 million donation to UC San Diego announced in March 2017 from UCSD Alumnus and Facebook co-founder Taner Halicioglu (Haw-li-dji-o-loo), several elements of SDSC's Data Science Hub framework will be integrated into the newly created Halicioglu Data Science Institute (HDSI). HDSI will focus on cross-disciplinary studies involving computer science, cognitive science, math and other fields. A formal launch of the institute is planned for early 2018.

"This generous gift will transform our institution and forever change the way we educate the next generation of scholars, which is what the Campaign for UC San Diego is about," said Chancellor Pradeep K. Khosla when the donation was announced by the university. "What we accomplish together in this campaign will lead to a future that is smarter, and brighter, than ever."



## 'AWESOME' SOCIAL MEDIA ANALYSIS

SDSC is pioneering an integrative analytics platform that harnesses state-of-the-art big data technologies to collect, analyze, and understand social media activity along with current events data and domain knowledge. Called AWESOME (for Analytical Workbench for Exploring SOCIAL Media) the platform is designed to help social science researchers, global health professionals, and government analysts use real-time, multi-lingual, citizen-level social media data to automatically crosslink it to relevant knowledge to determine significant issues that constitute common grounds, differences, polarization, and opportunities. This initiative has the potential to benefit society through areas as diverse as detecting free speech suppression, to shaping policy decisions, or even slowing the spread of viruses such as HIV or Zika.

A National Institutes of Health (NIH) award was recently granted to SDSC as well as UCLA and UC Irvine to use AWESOME begun in September 2017 to make timely predictions for the number of high HIV-risk patients at county levels. The award is an outcome of a collaboration between the groups under the banner of the UCOP-supported UC Institute for Prediction Technology (UCIPT), a multi-campus program to accelerate innovations in social technologies to predict human behaviors and outcomes. A second award, also begun in September 2017, was granted to UC San Diego and Princeton University from the NSF RIDIR (Resource Implementations for Data Intensive Research) program to perform multi-lingual integrative analysis of textual data from multiple sources such as newspapers, Twitter, Weibo, political biographies, etc., with the goal of identifying emerging situations that may require the attention of policy-makers and crisis managers.

AWESOME uses AsterixDB, a scalable data management system that can store, index, and manage semi-structured data that resulted from a research project at UC Irvine and UC Riverside. SDSC is collaboratively extending AsterixDB to enable new political and social science analytics for the 21st Century China Center initiative of UC San Diego, a project that produces scholarly research and informs policy discussions on China and U.S.-China relations.

SDSC's Centers of Excellence are part of a broad initiative to assist researchers across many data-intensive science domains, including those who are relatively new to computational and data-enabled science. These centers represent key elements of SDSC's wide range of expertise, from big data management to the analysis and advancement of the internet.



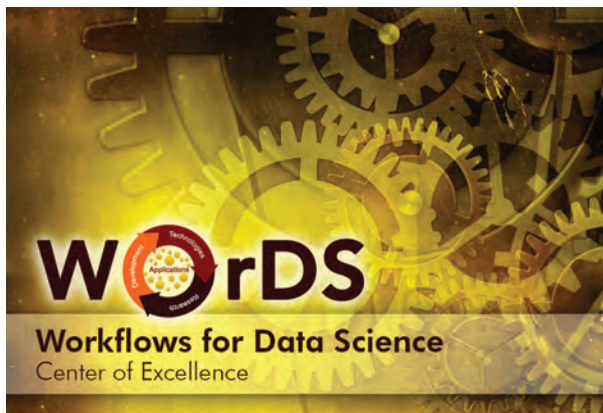
SDSC Chief Data Science Officer Ilkay Altintas also directs the WorDS Center and is principal investigator for the WIFIRE project, a university-wide collaboration funded by the National Science Foundation (NSF) to create a cyberinfrastructure to effectively monitor, predict, and mitigate wildfires.

## WORKFLOWS FOR DATA SCIENCE (WORDS) CENTER

Called WorDS for 'Workflows for Data Science', this center of excellence combines more than a decade of experience within SDSC's Scientific Workflow Automation Technologies laboratory, which developed and validated scientific workflows for researchers working in computational science, data science, and engineering. "Our goal with WorDS is to help researchers create their own workflows to better manage the tremendous amount of data being generated in so many scientific disciplines, while letting them focus on their specific areas of research instead of having to solve workflow issues and other computational challenges as their data analysis progresses from task to task," said Ilkay Altintas, director of WorDS and SDSC's Chief Data Science Officer.

Funded by a combination of sponsored agreements and recharge services. WorDS' expertise and services include:

- World-class researchers and developers well-versed in data science, big data and scientific computing technologies;
- Research on workflow management technologies that resulted in the collaborative development of the popular Kepler Scientific Workflow System;
- Development of data science workflow applications through a combination of tools, technologies, and best practices;
- Hands-on consulting on workflow technologies for big data and cloud systems, i.e., MapReduce, Hadoop, Yarn, Spark, and Flink; and
- Technology briefings and classes on end-to-end support for data science.



For more about WorDS use a QR code reader or visit <http://words.sdsc.edu>



Sandeep Chandra is director of SDSC's Health Cyberinfrastructure (CI) and head of the Sherlock center of excellence.

## SHERLOCK

Sherlock, an offering of SDSC's Health Cyberinfrastructure Division, is focused on managed information technology, compliance, and data services for academia and government that spans many IT disciplines, including compliant cloud hosting, cyber-security, and data management. Sherlock Cloud, a multi-tenant, scalable private Cloud, is compliant with the Federal Information System Management Act (FISMA), Health Information Portability and Accountability Act (HIPAA) and NIST 800-171 Controlled Unclassified Information (CUI) requirements.

Sherlock Cloud is collaborating with Amazon Web Services (AWS) to duplicate its managed services in the AWS Cloud, which would result in a hybrid cloud, offering customers the option to choose managed compliant services operating on site at SDSC or in the AWS Cloud. The collaboration is aimed at leveraging the scale and automation that such public clouds offer. The Health CI Division understands that public offerings such as AWS and Azure don't offer end-to-end compliant services; customers are expected to buy compliant compute and storage services and build the necessary management services on top to meet compliance. The Health CI Division's approach addresses this gap by building the necessary CI services on top of the public cloud resources, thereby offering customers end-to-end compliance.



For more about Sherlock  
use a QR code reader or visit  
<http://sherlock.sdsc.edu>

## CENTER FOR LARGE-SCALE DATA SYSTEMS RESEARCH

The CLDS research center was created at SDSC as a research partnership with industry to provide leadership for corporate sponsors and researchers interested in data- and technology-driven organizational transformation. CLDS researchers specialize in developing applicable concepts, frameworks, case analyses, and systems solutions to big data technology and technology management challenges. Current CLDS research areas include a multi-year effort to develop big data benchmarking, working with industry groups including the Transaction Processing Council (TPC) and the Standard Performance Evaluation Corporation (SPEC).

CLDS researchers were instrumental in developing BigBench, the industry's first end-to-end big data benchmark. CLDS research also measures the growth of data in the 'How Much Information?' research program, and the value of data in research on the asset value of organizational data. This work has been published in research journals and noted in business and general interest publications including *The Economist*, *The New York Times*, and the *Wall Street Journal*. As an industry/university partnership, CLDS encourages participation by industry researchers and analysts, and welcomes industry sponsorship of projects. Center research is available via a variety of venues, including working papers, research briefings, multi-company forum workshops, and an annual data-focused technology conference called DataWest.



For more about DataWest  
use a QR code reader or visit  
[www.datawest.org](http://www.datawest.org)



## Internet Research for Cybersecurity and Sustainability

The Center for Applied Internet Data Analysis (CAIDA) was the first center of excellence at SDSC. Formed in 1997, CAIDA is a commercial, government, and research collaboration aimed at promoting the engineering and maintenance of a robust and scalable global internet infrastructure. CAIDA's founder and director, KC Claffy, is a resident research scientist at SDSC whose research interests span internet topology, routing, security, economics, future internet architectures, and policy. Claffy was named 2017's recipient of the prestigious Jonathan B. Postel Service Award by the Internet Society, a global non-profit dedicated to ensuring the open development, evolution, and use of the internet. An adjunct professor of computer science and engineering at UC San Diego, Claffy has been with SDSC since 1991 and holds a Ph.D. in Computer Science from the university. The following are excerpts from a recent SDSC interview with Claffy.



**Q** What is CAIDA's mission today and has that mission changed since it was created in 1997?

**A** Our mission has evolved with the growth of the internet as a critical infrastructure, to better understand the security and stability of network infrastructure and architecture; develop and evaluate future information architectures, economics and public policy; and the ethics of information technology research.

**Q** Please provide an example of a recent CAIDA project that has a benefit for both science and society.

**A** CAIDA's recent appointment as the Independent Measurement Expert for the AT&T/DirecTV merger (a joint decision by AT&T and the U.S. Federal Communication Commission) is a gratifying signal that both the public and private sector respect CAIDA's expertise and objectivity enough to trust us with this important and unprecedented role in internet public policy.

**Q** All researchers share certain frustrations and/or barriers on their path to gratifying results. Can you describe one of the toughest research hurdles you've encountered?

**A** By far the biggest challenge is the opacity of the internet infrastructure, or in more operational terms, acquiring access to realistic and representative data sets, and validation of inferences from our own measurements and models.

**Q** How can we help the public understand the complex nature of the internet so we can make more informed decisions about its future?

**A** Inspired by the SDSC Science Gateway Institute, we recently proposed a large collaboration to develop a new science gateway that will offer interdisciplinary researchers more accessible, calibrated and user-friendly tools to collect, analyze, query, and interpret measurements of the internet ecosystem.



For more information about CAIDA use a QR code reader or visit <https://goo.gl/JTDCik>





More information on WIFIRE at  
<https://wifire.ucsd.edu>

## Los Angeles Fire Department, UC San Diego WIFIRE Team Join Forces to Fight Wildfires

In late 2016, the Los Angeles Fire Department, challenged by yet another series of late summer wildfires, successfully tested a new web-based tool developed by UC San Diego researchers to perform data-driven predictive modeling and analysis of fires that have a high potential for rapid spread.

With 2017 being another year of wind-whipped wildfire activity into December, the LAFD and other departments and agencies relied on the new tool, called Firemap and developed by UC San Diego's 'WIFIRE' collaboration that provides a 'what-if' analysis of fire scenarios ahead of time as well as real-time fire forecasting. During the devastating wildfires in Santa Rosa in October 2017, the part of the WIFIRE tool that can be accessed by the public generated more than 1.5 million views as residents monitored the outbreak that destroyed some 3,500 homes and businesses.

The overall goal of WIFIRE, the result of a multi-year National Science Foundation (NSF) grant, is to make data and predictive models readily available so that the direction and rate of fire spread can be known as early as possible to assist in rescue and containment efforts.

WIFIRE's Firemap data resource also provides easy access to information on past fires, past and current weather conditions as well as weather forecasts, satellite detections as fast as they are received, images from an extensive network of internet-based cameras operated by the High-Performance Wireless Research & Education Network (HPWREN), and information on vegetation and landscapes from a variety of sources. These datasets are being used for planning fire response and management of natural resources well ahead of time.

"WIFIRE is a great example of how distributed data sets can be used within an integrated system to transform that information to action in real-time for an effective decision support and response," said Ilkay Altintas, SDSC's chief data science officer and the principal investigator for WIFIRE.

"For Incident Commanders (IC), the WIFIRE Firemap is one of the most progressive decision-making tools developed in the last decade," said LAFD Fire Chief Ralph Terrazas. "Firemap gives the IC accurate and real-time data to help make command decisions when prioritizing resource allocation or which communities to evacuate. This has never been available during the initial action phase of brush firefighting, and it has been an honor to work with the WIFIRE team and see their dedication to public safety." Please read more about HPWREN's role in fighting wildfires on Page 34.



## INDUSTRIAL RELATIONS



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 continued to be recognized by industry during the 2016/17 fiscal year for its thought leadership and expertise across a broad range of emerging technologies, including the Internet of Things (IoT), smart manufacturing, machine learning and artificial intelligence, life sciences computing, high-performance computing, and all things data.

### INTERNET OF THINGS

During 2016, the Internet of Things (IoT) – the rapidly growing network of physical objects that contain embedded computing technology to communicate and sense or interact with their internal states or the external environment – emerged as a theme in relation to both smart manufacturing and smart grid technologies. A study published in 2015 by Deloitte Touche Tohmatsu found that “advanced manufacturing industries” accounted for 70% of the 40 million jobs and \$2.7 trillion in economic output generated by in the U.S. The study further found that the IoT and Predictive Data Analytics were two of the most promising technologies for driving innovation and competitiveness in U.S. advanced manufacturing activities.

IoT and predictive analytics can work hand-in-hand to optimize advanced manufacturing systems, leading to increased efficiencies, throughput, and quality control. IoT permits the deployment of thousands of low-cost sensors throughout a manufacturing line and even on the manufactured components themselves. Continuous measurements can be streamed and collected in a “data lake,” where predictive analytics and machine learning techniques can be applied to reveal hidden insights on how to improve the processes.



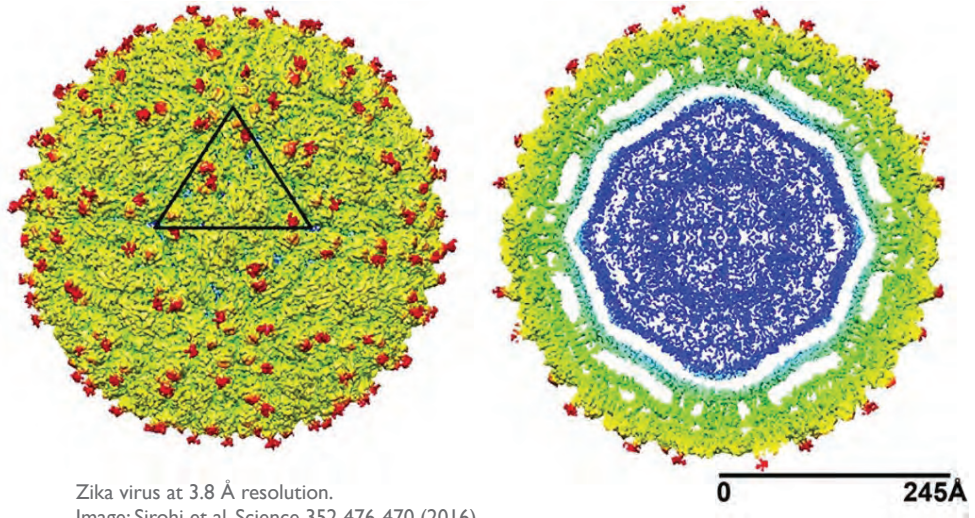




Bob Sinkovits is Director of Scientific Computing Applications at SDSC. He has collaborated with researchers spanning many fields including physics, chemistry, astronomy, structural biology, finance, ecology, climate, immunology and the social sciences, with an emphasis on making the most effective use of high-end computing resources.

## LIFE SCIENCES COMPUTING

Technological advances in DNA sequencing and cryo-electron microscopy are generating vast amounts of data that are enabling research discoveries crucial to understanding illnesses and developing personalized medical treatments. Analyzing these data requires continual development of better and faster computational techniques. In late 2016, SDSC, with funding from Dell and Intel under the Dell “Centers for Innovation” program, carried out a project to characterize the performance and identify optimization opportunities for key bioinformatics applications on high-performance computing systems. Selected applications included “GATK,” a “gold standard” genome analysis toolkit developed by the Broad Institute, and “RELION,” an emerging *de facto* standard application for analyzing cryo-electron microscopy data. Through this study, SDSC researchers developed key insights into HPC systems characteristics that influence the performance of bioinformatics applications, leading to valuable guidance to both academic and commercial researchers for purchasing and configuring computing and storage systems for life sciences research.



Zika virus at 3.8 Å resolution.  
Image: Sirohi et al. Science 352 476-470 (2016)

## MACHINE LEARNING & ARTIFICIAL INTELLIGENCE

Interest in applications of machine learning and artificial intelligence (AI) continued to gain momentum during this period and fostered engagement with industry. SDSC hosted several AI-related meetups and mixers sponsored by the Machine Learning Society and Analytics Ventures, an investment firm focused on AI, machine learning, and the IoT. The robust attendance and energy level at these events suggested current and would-be technology entrepreneurs in San Diego are enthusiastic about building technologies and businesses around AI. SDSC scientists, with established expertise in this field, are conducting research and developing training for AI, analytics, and machine learning for academia and commerce in areas as diverse as autonomous vehicles, cancer treatment, and firefighting, among others.

For example, the Center is delivering global online training in big data and machine learning via Coursera and EdX. A recent NSF-funded supplement to augment SDSC’s *Comet* supercomputer with additional NVIDIA Graphics Processing Units (GPUs) increases that system’s capability for training deep-learning networks for both scientific and industrial applications.



## EMERGING OPPORTUNITIES IN DATA

Since its beginning, SDSC's industrial programs have sought to provide a neutral forum where companies can come together to explore emerging technology topics of mutual interest. In this spirit, in December of 2016, SDSC held the inaugural "Data West" conference. The aim of the conference was to provide a discussion-oriented, information exchange forum on emerging business opportunities in data geared towards senior thought leaders and decision-makers. Organized by SDSC Principal Investigator Jim Short and his collaborator, Lynda Applegate of the Harvard Business School, Data West brought together a diverse group of about 100 attendees from large and small companies; government representatives from the Department of the Navy and City of San Diego; and scientists from UC San Diego, MIT, and elsewhere.

Data West sponsors – Intel, Seagate, Dell EMC, Mellanox, Back Bay Data Solutions, Booz Allen Hamilton, Cray, Collibra, Ampool, Peaxy, VelociData, and First San Francisco Partners – provided both attendees and generous support for the event. Over the course of two days, the participants engaged in lively and wide-ranging discussions on topics such as large-scale data system agility and performance, new ventures in data analytics, cyber-physical systems, machine learning, and smart manufacturing.



For more on the Industry Partners Program use a QR code reader (left) or visit [www.sdsc.edu/collaborate/ipp.html](http://www.sdsc.edu/collaborate/ipp.html) and [www.datawest.org](http://www.datawest.org) (right QR)



# FACTS & FIGURES

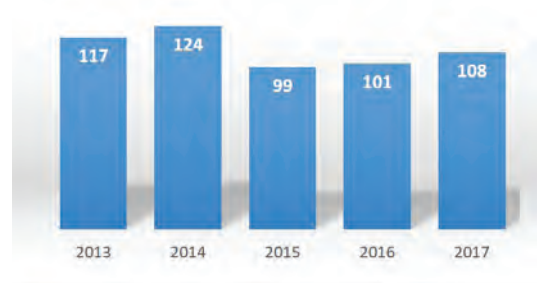
## PROPOSAL SUCCESS RATE

	FY13	FY14	FY15	FY16	FY17
Proposals Submitted	84	77	91	73	84
Proposals Funded	37	39	38	33	33
Success Rate	45%	50%	42%	45%	39%

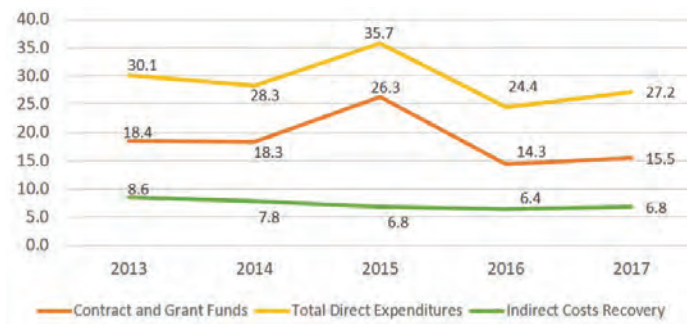
(Source: Campus Data Report)

In perhaps the most competitive landscape for federal funding in the last two decades, SDSC's overall success rate on federal proposals averages 44% over the last five years compared to the 2017 national average of about 18% for computer science and engineering proposals at the National Science Foundation.

## NUMBER OF SPONSORED RESEARCH AWARDS



## TOTAL EXPENDITURES (\$M)

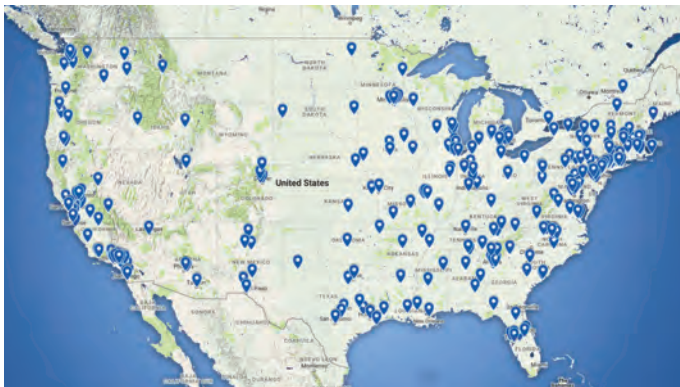


Apart from the extraordinary research impact of SDSC collaborations and partnerships, a quick look at the fiscal impact of these collaborations is impressive. During its 32-year history, SDSC revenues have exceeded \$1 billion, a level of sustained funding matched by few academic research units in the country. At the close of the 2017 fiscal year, SDSC had 59 NSF-funded projects totaling \$127 million.

## TOTAL REVENUE FROM INDUSTRY (\$M)



## GEOGRAPHICAL DISTRIBUTION OF NATIONAL USERS OF SDSC HPC RESOURCES

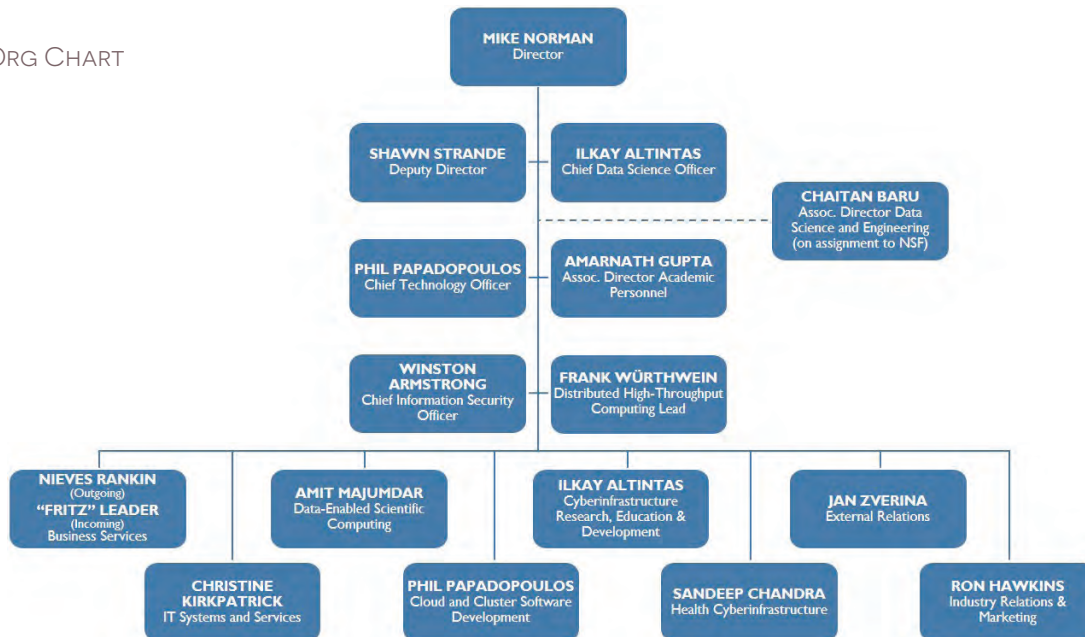


A total of 2,549 unique users from around the world accessed SDSC's *Comet* HPC resource directly. Of these users, 2,489 were based in the United States. The map to the left illustrates a sampling of the U.S. locations of these users. In addition more than 12,230 unique users who did research via science gateways during FY2016-17.

On *Comet*, a total of 385,803,699 service units (SUs) were used during the fiscal period.

# ORGANIZATION & LEADERSHIP

SDSC ORG CHART



## EXECUTIVE TEAM

- Ilkay Altintas**  
Chief Data Science Officer
- Chaitanya Baru** (on assignment to NSF)  
Associate Director, Data Science and Engineering
- Sandeep Chandra**  
Division Director, Health Cyberinfrastructure
- Ronald Hawkins**  
Director, Industry Relations
- Christine Kirkpatrick**  
Division Director, IT Systems and Services
- Samuel "Fritz" Leader** (incoming)  
Chief Administrative Officer
- Amit Majumdar**  
Division Director, Data-Enabled Scientific Computing
- Michael Norman**  
SDSC Director
- Philip M. Papadopoulos**  
Division Director, Cloud & Cluster Software Development
- Nieves Rankin** (outgoing)  
Division Director, Business Services
- Shawn Strande**  
Deputy Director
- Frank Würthwein**  
Lead, Distributed High-Throughput Computing
- Jan Zverina**  
Division Director, External Relations

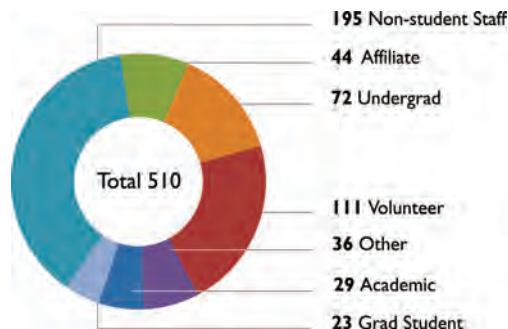
## UC EXTERNAL ADVISORY BOARD

- Kimberly S. Budil**  
UC, Office of the President (UCOP)
- Michael Carey**  
UC Irvine
- Trish Damkroger**  
Lawrence Livermore National Laboratory
- Paul Dodd**  
UC Davis
- Adams Dudley**  
UC San Francisco
- Lise Getoor**  
UC Santa Cruz
- Ralph Greenspan**  
UC San Diego
- Eamonn Keogh**  
UC Riverside
- Juan C. Meza**  
UC Merced
- Peter Nugent**  
Lawrence Berkeley National Laboratory and UC Berkeley
- John Sarrao**  
Los Alamos National Laboratory
- Tim Sherwood**  
UC Santa Barbara
- Paul Weiss**  
UCLA
- Tarek I. Zohdi**  
UC Berkeley

## EXECUTIVE COMMITTEE

- |                     |                        |
|---------------------|------------------------|
| UC SAN DIEGO        | SDSC                   |
| Sandra Brown        | Ilkay Altintas         |
| Mark Ellisman       | Chaitanya Baru         |
| Michael Holst       | Sandeep Chandra        |
| J. Andrew McCammon  | Ronald Hawkins         |
| John Orcutt         | Christine Kirkpatrick  |
| Al Pisano (chair)   | Amit Majumdar          |
| Tajana Rosing       | Michael Norman         |
| Nicholas Schork     | Philip M. Papadopoulos |
| Brian Schottlaender | Nieves Rankin          |
| Robert Sullivan     | Shawn Strande          |
| Susan Taylor        | Frank Würthwein        |
| Gabriel Wienhausen  | Jan Zverina            |

## SDSC CENSUS FY2016



# RESEARCH EXPERTS

## SDSC COMPUTATIONAL & DATA SCIENTISTS

### **Ilkay Altintas, Ph.D.**

*Chief Data Science Officer, SDSC*  
*Director, Workflows for Data Science (WorDS) Center of Excellence*  
*Lecturer, Computer Science and Engineering, UCSD*  
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

### **Chaitan Baru, Ph.D.**

*SDSC Distinguished Scientist*  
*Director, Center for Large-scale Data Systems Research (CLDS), SDSC*  
*Associate Director, Data Science and Engineering, SDSC*  
*Associate Director, Data Initiatives, SDSC*  
Data management  
Large-scale data systems  
Data analytics  
Parallel database systems

### **Hans-Werner Braun, Ph.D.**

*Research Scientist Emeritus, SDSC*  
Internet infrastructure, measurement/analysis tools  
Wireless and sensor networks  
Internet pioneer (PI, NSFNET backbone project)  
Multi-disciplinary and multi-institutional collaborations

### **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

### **Sandeep Chandra, M.S.**

*Executive Director, Sherlock Cloud*  
*Director, Health Cyberinfrastructure Division, SDSC*  
Compliance (NIST, FISMA, HIPAA)  
Scientific Data Management  
Cloud Computing  
Systems Architecture & Infrastructure Management

### **Dong Ju Choi, Ph.D.**

*Senior Computational Scientist, SDSC*  
HPC software, programming, optimization  
Visualization  
Database and web programming  
Finite element analysis

### **Amit Chourasia, M.S.**

*Senior Visualization Scientist, SDSC*  
*Lead, Visualization Group*  
*Principal Investigator, SEEDME.org*  
Visualization and computer graphics  
Ubiquitous Sharing Infrastructure

### **Pietro Cicotti, Ph.D.**

*Senior Computational Scientist, SDSC*  
Architecture  
Runtime systems  
Performance tools  
Analysis, optimization, and modeling

### **KC Claffy, Ph.D.**

*Director/PI, CAIDA (Center for Applied 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

### **Yifeng Cui, Ph.D.**

*Director, Intel Parallel Computing Center, SDSC*  
*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

### **Alberto Dainotti, Ph.D.**

*Assistant Research Scientist, CAIDA (Center for Applied Internet Data Analysis)*  
Internet measurements  
Traffic analysis  
Network security  
Large-scale internet events

### **Amogh Dhamdhere, Ph.D.**

*Assistant Research Scientist, CAIDA (Center for Applied Internet Data Analysis)*  
Internet topology and traffic  
Internet economics  
IPv6 topology and performance  
Network monitoring and troubleshooting

### **Andreas Goetz, Ph.D.**

*Co-Director, CUDA Teaching Center*  
*Co-Principal Investigator, Intel Parallel Computing Center*  
Quantum Chemistry  
Molecular Dynamics  
ADF and AMBER developer  
GPU accelerated computing

### **Amarnath Gupta, Ph.D.**

*Associate Director, Academic Personnel, SDSC*  
*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

### **Amit Majumdar, Ph.D.**

*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

**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

**Michael Norman, Ph.D.**

*Director, San Diego Supercomputer Center*  
*Distinguished Professor, Physics, UCSD*  
*Director, Laboratory for Computational Astrophysics, UCSD*  
Computational astrophysics

**Francesco Paesani, Ph.D.**

*Lead, Laboratory for Theoretical and Computational Chemistry, UCSD*  
Theoretical chemistry  
Computational chemistry  
Physical chemistry

**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/OptlPuter  
Prism@UCSD

**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

**Andreas Prlić, Ph.D.**

*Technical and Scientific Team Lead, RCSB Protein Data Bank*  
Bioinformatics  
Structural biology  
Computational biology  
Protein Data Bank

**Peter Rose, Ph.D.**

*Director – Structural Bioinformatics Laboratory*  
*Lead – Bioinformatics and Biomedical Applications, Data Science Office*  
Structure-based drug design  
Bioinformatics  
Computational biology  
Protein Data Bank

**Robert Sinkovits, Ph.D.**

*Director, Scientific Computing Applications, SDSC*  
High-performance computing  
Software optimization and parallelization  
Structural biology  
Bioinformatics  
Immunology  
Relational databases

**Britton Smith, Ph.D.**

*Assistant Research Scientist, SDSC*  
Computational astrophysics  
Software development

**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

**David Valentine, Ph.D.**

*Research Programmer, Spatial Information Systems Laboratory, SDSC*  
Spatial and temporal data integration/analysis  
Geographic information systems  
Hydrology  
Spatial management infrastructure

**Nancy Wilkins-Diehr, M.S.**

*Associate Director, SDSC*  
*Co-Principal Director, XSEDE at SDSC*  
*Co-Director for Extended Collaborative Support, XSEDE*  
Science gateways  
User services  
Aerospace engineering

**Frank Würthwein, Ph.D.**

*Distributed High-Throughput Computing Lead, SDSC*  
*Executive Director, Open Science Grid*  
*Professor of Physics, UCSD*  
High-capacity Data Cyberinfrastructure  
High-energy Particle Physics

**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](http://www.sdsc.edu)

[twitter/SDSC\\_UCSD](https://twitter.com/SDSC_UCSD)

[facebook/SanDiegoSupercomputerCenter](https://facebook.com/SanDiegoSupercomputerCenter)