PARTNERING
for Resources and Know-how

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 INDUSTRY PARTNERS PROGRAM

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

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

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

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

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

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

CLDS AND PACE: 'BIG DATA' CENTERS OF EXCELLENCE

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

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

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

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

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

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

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

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

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

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

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

Courtesy of Hunter Industries

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

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

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

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

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

SPEEDING BIG DATA GENOMIC ANALYSIS OF RHEUMATOID ARTHRITIS

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

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

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

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

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

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

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

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

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

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

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

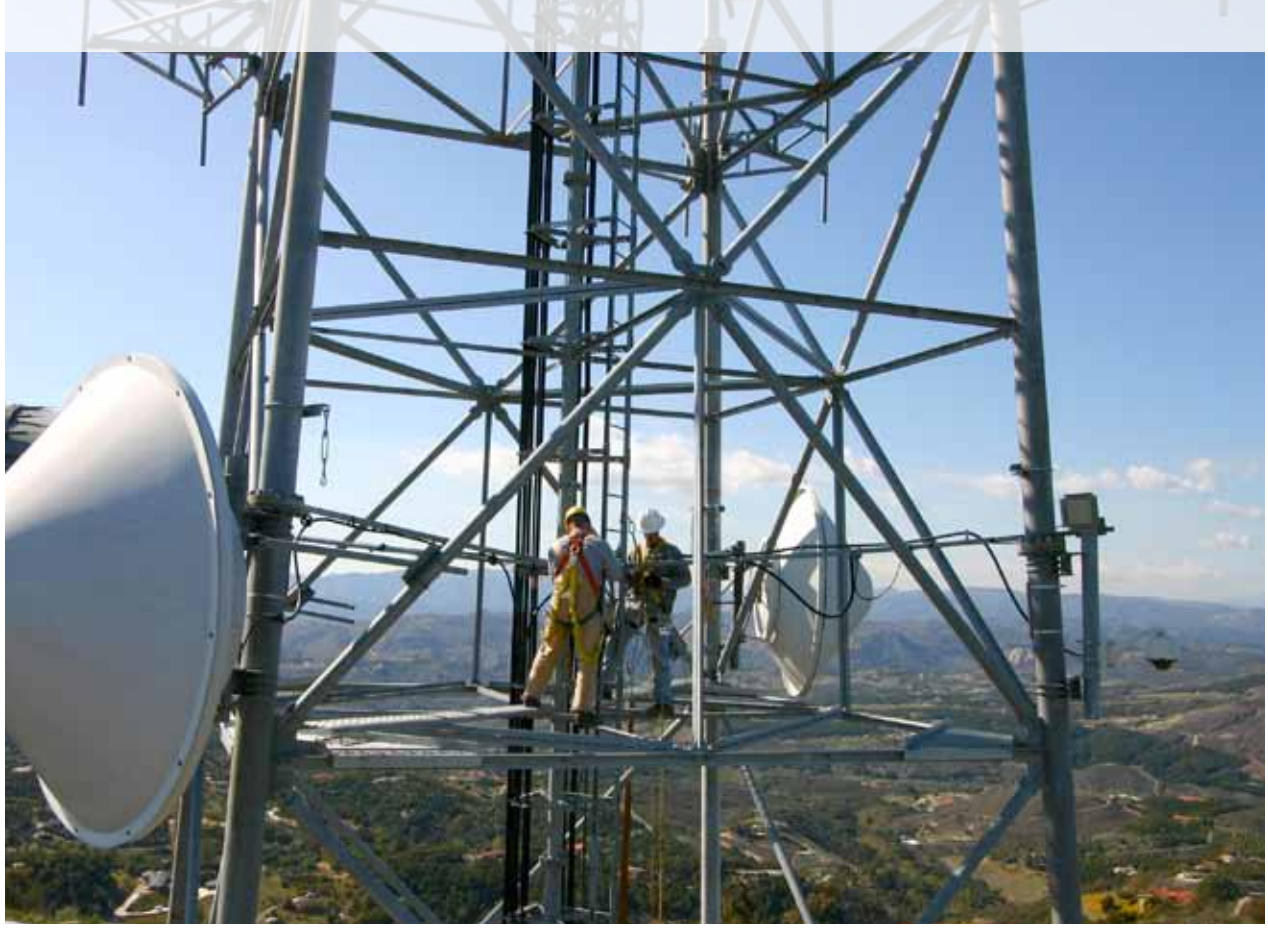
When a fast-moving blaze burned more than 7,000 acres near Mount Laguna in August 2013, firefighters were able to monitor its spread and respond accordingly by relying on a high-speed data transmission network made possible in part by UC San Diego's High-Performance Wireless and Research Education Network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

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

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

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

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



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

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

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

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

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

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

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