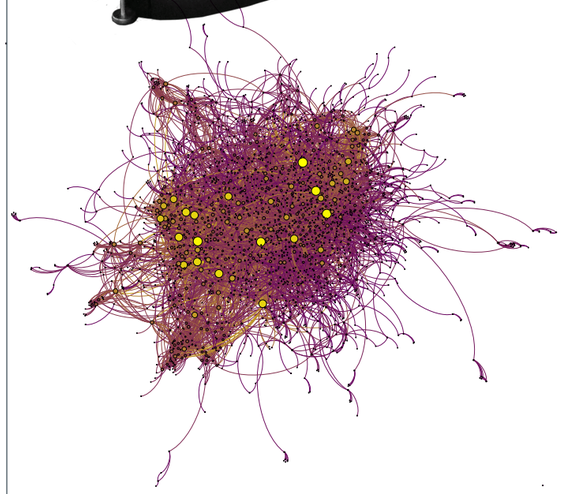
Sherlock: Unlocking the Secrets of Big Data

A YarcData uRiKA system representative of PSC's Sherlock

Computational analysis that discovers underlying patterns in "big data" can open many doors to understanding, such as how genes work, the dynamics of social networks, and the source of breaches in computer security. With this kind of analysis, based on a mathematical approach called "graph theory," interconnected webs of information can be represented as graphs, wherein nodes represent data elements and edges represent relationships among them.

Such graphs produced from real-world data can be huge, containing billions or trillions of edges. Even more challenging, these graphs typically can't be partitioned; their high connectivity prevents dividing them into subgraphs that can be practically mapped onto distributed-memory computers. "Graph analytics are notoriously difficult," says Nick Nystrom, PSC's director of strategic applications, "because following unpredictable paths from node-to-node is rate-limited by latencies to remote and local memory, which has drastically limited the graph problems that can be tackled."

To break the barrier blocking large-scale graph analytics, PSC this year introduced Sherlock, a unique supercomputer specialized for complex analytics on big data, which will be used for pilot projects by the national research community.



Protein-protein interactions in yeast, forming a relatively small graph of only 7,182 edges, illustrate the complexity of problems in graph analytics. (See Vladimir Batagelj & Andrej Mrvar (2006): Pajek datasets, http://vlado.fmf.uni-lj.si/pub/networks/data)

Sherlock: The Details

Acquired through NSF's Strategic Technologies for Cyberinfrastructure program, Sherlock is a YarcData uRiKA ("Universal RDF Integration Knowledge Appliance") data appliance. It features massive multi-threading, shared memory, and hardware optimizations to enable exceptionally efficient execution of graph algorithms. Sherlock contains 32 next-generation Cray XMT nodes. Aggregate shared memory is one terabyte, which can accommodate a graph of approximately 10 billion edges.

PSC customized Sherlock via additional Cray XT5 nodes having AMD Opteron processors to add valuable support for heterogeneous applications that use the XMT nodes as accelerators for graph-based algorithms. This heterogeneous capability will enable an even broader class of applications, for example in genomics, astrophysics, and other types of analysis of complex networks.

Sherlock runs an enhanced suite of familiar semantic web software for easy access to powerful analytic functionality, using the Resource Description Framework (RDF) as a very general and expressive data format. Sherlock also supports common programming languages such as C, C++, Java, Fortran, and scripting languages.

Supercomputing in Pennsylvania

PSC provides education, consulting, advanced network access and computational resources to scientists and engineers, teachers and students across the Commonwealth of Pennsylvania

3ROX:Network for Education

The Three Rivers Optical Exchange (3ROX) (see pp. 12-13) provides research and education network service to seven Intermediate Units in western Pennsylvania that serve 116 school districts, more than 600 schools, 21,000 teachers and 300,000 students. 3ROX links these schools, teachers and students to a global community of people and ideas.

Research & Training in Pennsylvania

Researchers in Pennsylvania are using PSC's new disk-based data storage, the Data Supercell (see p. 4), implemented with support from the Commonwealth of Pennsylvania's Redevelopment Assistance Capital Program. Pennsylvania organizations using the Data Supercell include the National Energy Technology Laboratory, Carnegie Mellon University, the Software Engineering Institute, the University of Pittsburgh Developmental Biology Program, and Drexel University's Design Arts Group.

Continuing a long-standing relationship with Lehigh University, PSC in March did a half-day workshop on parallel programming of multi-core computing systems. PSC scientist John Urbanic presented material on programmer-friendly standards (OpenMP and Open ACC) to 35 students as part of Lehigh's annual HPC Symposium.

In June, PSC scientific co-director Ralph Roskies and PSC director of networking Wendy Huntoon addressed information officers and other leaders of the 14 universities of the Pennsylvania State System of Higher Education. Their presentation highlighted how computational science and cyberinsfrastructure are changing research and outlined possible collaboration between PSC and PASSHE universities.

As part of a program sponsored by PAIUnet, Pennsylvania's statewide, high-speed educational network, PSC in March helped to develop and coordinate a data-modeling session for high-school students taking part in the state-wide Marcellus Shale Project. Also, through its BEST and CAST programs (see pp. 8-9), PSC provides continuing training and curriculum materials for western Pennsylvania high-school math and science teachers.

Pennsylvania Research Innovation

Several projects in this booklet highlight research in Pennsylvania enabled through PSC:

- **Bright Lights, Big Cosmos:** Astrophysicists at Carnegie Mellon University are simulating the period in the evolution of the Universe when stars, galaxies and black holes first appeared (p. 40).
- **Modeling Aortic Aneurysms:** Drawing on data from Allegheny General Hospital, biomedical engineers are modeling aortic aneurysms so that it will be possible to better guide decisions on when surgery is required (p. 44).
- **When Small Worlds Collide:** A Lehigh University physicist is calculating spin properties of molecules that could help lead to quantum computing, much faster than today's supercomputers (p. 45).
- **Fighting Dengue Resurgence:** Researchers at PSC and the University of Pittsburgh are developing tools to help public-health decision makers intervene effectively to stop the world-wide spread of dengue fever (p. 47).

Shared Memory Poker

Using PSC's Blacklight system, Carnegie Mellon computer science professor Tuomas Sandholm and his Ph.D. student Sam Ganzfried did well at Toronto in July — the Advancement of Artificial Intelligence (AI) annual Computer Poker Competition. In recent years, poker has emerged as an AI challenge similar to that served for many years by chess, but more demanding. "In poker," says Sandholm, "a player doesn't know which cards the other player holds or what cards will be dealt in the future. Such games of incomplete information are much harder to solve than complete-information games."

Sandholm's field, game theory, in which his work is internationally recognized, describes conflict in which the payoff is affected by actions and counter-actions of intelligent opponents. Like many games, poker can be formulated mathematically, but the formulations are unimaginably huge. Two-player no-limit Texas Hold'em poker has a "game tree" of about 10^{71} nodes, hence the usefulness of large amounts of memory. At Toronto, running with Blacklight, the Sandholm group's poker-playing agent finished second in the instant runoff scoring for two-player no-limit Texas Hold'em.

Supercomputing Provided to Pennsylvania Organizations

From July 2011 through June 2012, PSC provided more than 7.8 million processor hours to 917 individual Pennsylvania researchers from 40 institutions. The following Pennsylvania corporations, universities, colleges and K-12 institutions used PSC resources during this period:

Albright College
Allegheny General Hospital
Allegheny-Singer Research Institute
Bryn Mawr College
Carnegie Mellon University
Cedar Crest College
Cheyney University of Pennsylvania
Community College of Allegheny County
Dickinson College
Drexel University
Duquesne University
Dynamix Technologies
Frazier School District
Grove City College
Haverford College
Indiana University of PA, all campuses
Kutztown University of Pennsylvania
Lehigh University
Life Technologies
Lock Haven University
Marconi Services
Oakland Catholic High School
Our Lady of Sacred Heart High School
PA CYBER Charter School
Pennsylvania State University, all campuses
Pittsburgh Public Schools
Pittsburgh Supercomputing Center
Shippensburg University of Pennsylvania
Slippery Rock University
Swarthmore College
Temple University
Thomas Jefferson University
University of Pennsylvania
University of Pittsburgh, all campuses
University of the Sciences in Philadelphia
Upper St. Clair High School
Ursinus College
Vitaerx
Wilkes University
Winchester-Thurston School