Institute for **DATA** and **HIGH PERFORMANCE COMPUTING**

*Leadership in Scientific Discovery*
Big data and high performance computing (HPC) are playing essential roles in attacking the most important problems that face society today. The Institute for Data and HPC (IDH) provides a fertile environment and an organization to create and advance innovative, interdisciplinary research efforts in these areas. Georgia Tech faculty make up a robust community of researchers spanning a broad range of key areas including applications, algorithms, hardware and software systems, and the underlying foundational mathematics.

We are proud of Georgia Tech’s reputation and track record for interdisciplinary research and our ability to develop and apply technology toward the solution of science and engineering’s toughest problems.

I invite you to learn more about our talented researchers and exciting research and education programs by reviewing this brochure and visiting our website at www.idh.gatech.edu. Please don’t hesitate to contact us for additional information.

Sincerely,

Richard Fujimoto
Interim Director, Institute for Data and HPC
Regents’ Professor and Chair, School of Computational Science & Engineering
Georgia Institute of Technology

About Institute For Data & High Performance Computing

High performance computing is about much more than speed—it’s about what you can achieve with extreme capability. At Georgia Tech, we don’t just have people and resources; we have the right people and the right resources to lead computation-driven scientific discovery. IDH leverages investments and research in data intensive and high performance computing at Georgia Tech for the benefit of our modern society. We do this by creating and strengthening multidisciplinary research teams that combine deep knowledge of HPC application areas with advances in computational techniques and foundational mathematics to attack the most challenging problems facing society in science, engineering, and the social sciences. From experimental systems to applied research, we are addressing the world’s most important challenges, including:

- **Healthcare** — diagnosis of heart disease and stroke, identification of tumors, detection and prevention of epidemics and pandemics
- **Bioinformatics and Systems Biology** — protein folding, drug design, biological system simulation, complex life systems, microbial research, HIV virus, life and evolution
- **Sustainability and Urban Infrastructure** — city planning, intelligent transportation systems, communication, water, food supply, emergency planning
- **Energy** — smart electric power grids, combustion, fusion, energy conservation behavior
- **Nanoscience and Nanotechnology** — nanomedicine, manufacturing, electronics and supercomputers, consumer products
- **Cybersecurity** — Web science and interaction networks, Internet security, social networks
- **Data-Intensive Analytics** — business analytics, streaming graph problems, national security
- **Homeland Security** — text analysis, fingerprint and face recognition, data and visual analytics

About the School of Computational Science & Engineering

Georgia Tech is devoted to the advancement and promotion of Computational Science & Engineering (CSE). Computational Science and Engineering is a discipline devoted to the systematic study, creation, and application of computer-based models to understand and analyze natural and engineered systems. We believe the CSE discipline fundamentally derives much of its richness and potential for impact from collaboration with other disciplines. Thus, the school has a strong emphasis on interdisciplinary research and education.

CSE research spans many computational areas. For example, research in high performance computing develops new ways to exploit the world’s most powerful supercomputers. Research in massive scale data and visual analytics and machine learning explores ways to extract useful information from the unprecedented volumes of data now appearing on the Internet and in many fields of science, engineering, and medicine. Modeling and simulation research explores new methods to exploit parallel and distributed computing platforms in order to solve challenging problems in areas such as medicine and transportation. Algorithm research builds a solid foundation spanning both continuous and discrete models. Our research includes interdisciplinary collaborations and interactions that crisscross the Georgia Tech campus—and extend around the world.

Computational Science and Engineering graduate degrees are jointly offered by the Colleges of Computing, Engineering, and Sciences at Georgia Tech.

- MS in Computational Science & Engineering (including a distance learning option)
- PhD in Computational Science & Engineering
University of Houston. Team members from Georgia Tech include Sudhakar Yalamanchili, Greg Eisenhauer and Matthew Wolf.
The team leads are Karsten Schwan at Georgia Tech; Allen Malony at the University of Oregon; and Barbara Chapman at the
University of Houston. Team members from Georgia Tech include Sudhakar Yalamanchili, Greg Eisenhauer and Matthew Wolf.

The team leads are Karsten Schwan at Georgia Tech; Allen Malony at the University of Oregon; and Barbara Chapman at the
University of Houston. Team members from Georgia Tech include Sudhakar Yalamanchili, Greg Eisenhauer and Matthew Wolf.

An improved understanding of turbulence will have great significance for both fundamental science and a wide range of applications in nature and engineering.

To that end, P. K. Yeung is leading a team of researchers in simulating turbulent mixing and dispersion at a record-setting resolution of 64 billion grid points. These simulations track the motions of large numbers of small particles in fluids and closely examine their movement away from one another under the influence of differing lengths and time scales, providing unique insights into fluid physics and aiding model development.

The science impacts of these simulations will include new models of pollutant dispersion based on a rigorous understanding of turbulence-flow physics; the modeling of heat and mass transport in combustion, industrial processing, and the environment; and new insights into oceanic (and atmospheric) mixing and dispersion phenomena, which play an important role in global climate change.

The team is striving to simulate the fluctuations in temperature and molecular concentrations of the infinitesimal fluid elements in turbulence using parameters typical in air and water. In addition to reaching a high Reynolds number, the team will also tackle the special challenges of simulating mixing in parameter regimes in which molecular diffusion is slow compared to the diffusion of momentum. Understanding mixing is especially important in stratified flow such as that found in the ocean.

The project explores the information that can be shared across the software stack used in HEC machines. The methods being developed analyze program information, performance data and tool knowledge. The resulting Glass Box system will allow developers to better assess the performance of their parallel codes. Tool creators can use the performance data to create new analysis and optimization techniques. System developers can also better manage multicores and machine resources at runtime, using JIT compilation and binary code editing to exploit the evolving hardware. Working with the Keeneland’ NSF Track II machine and industry partners, the project is creating new performance monitoring tools, compiler methods and system-level resource management techniques. The effort is driven by the large-scale codes running on today’s petascale machines. Its broader impact is derived from the interactions with technology developers and application scientists as well as from its base in three universities with diverse student populations.

The team leads are Karsten Schwan at Georgia Tech; Allen Malony at the University of Oregon; and Barbara Chapman at the
University of Houston. Team members from Georgia Tech include Sudhakar Yalamanchili, Greg Eisenhauer and Matthew Wolf.

An improved understanding of turbulence will have great significance for both fundamental science and a wide range of applications in nature and engineering.

To that end, P. K. Yeung is leading a team of researchers in simulating turbulent mixing and dispersion at a record-setting resolution of 64 billion grid points. These simulations track the motions of large numbers of small particles in fluids and closely examine their movement away from one another under the influence of differing lengths and time scales, providing unique insights into fluid physics and aiding model development.

The science impacts of these simulations will include new models of pollutant dispersion based on a rigorous understanding of turbulence-flow physics; the modeling of heat and mass transport in combustion, industrial processing, and the environment; and new insights into oceanic (and atmospheric) mixing and dispersion phenomena, which play an important role in global climate change.

The team is striving to simulate the fluctuations in temperature and molecular concentrations of the infinitesimal fluid elements in turbulence using parameters typical in air and water. In addition to reaching a high Reynolds number, the team will also tackle the special challenges of simulating mixing in parameter regimes in which molecular diffusion is slow compared to the diffusion of momentum. Understanding mixing is especially important in stratified flow such as that found in the ocean.
Computational Quantum Chemistry

In the past, quantum chemistry methods have not scaled very well with the size of the molecular system because of limitations in developing and implementing efficient parallel algorithms. The PSI4 Project, led by David Sherrill, takes a giant step forward in the quantum chemistry field by filling this need. In combining the density-fitting and Cholesky decomposition with localized-electron approximations, their research is dramatically reducing the computational costs associated with the complexities of explaining and predicting chemical behaviors.

By leveraging the very popular open-source quantum chemistry program (PSI3), Georgia Tech continues their tradition of developing key numerical techniques and parallel algorithms by focusing on enabling some of the fastest implementations of widely-used functions. In some preliminary analysis, Sherrill’s group was able to produce accurate computations in 200 hours of computer time on a single 8-core machine with 64 GB of memory, whereas the same computation previously took over 6000 hours for similar accuracy. These types of increases are strong indicators that massively-parallel implementations of these algorithms will be the next state-of-the-art in computational quantum chemistry.

Massively-parallel implementations of these algorithms will be the next state-of-the-art in computational quantum chemistry.

Georgia Tech Develops Computational Algorithm to Assist in Cancer Treatments

New Software Key for Personalized Cancer Medicine

High-throughput DNA sequencing technologies are leading to a revolution in how clinicians diagnose and treat cancer. The molecular profiles of individual tumors are beginning to be used in the design of chemotherapeutic programs optimized for the treatment of individual patients. The real revolution, however, is coming with the emerging capability to inexpensively and accurately sequence the entire genome of cancers, allowing for the identification of specific mutations responsible for the disease in individual patients.

There is only one downside. Those sequencing technologies provide massive amounts of data that are not easily processed and translated by scientists. That’s why Georgia Tech has created a new data analysis algorithm that quickly transforms complex RNA sequence data into usable content for biologists and clinicians. The RNA-Seq analysis pipeline (R-SAP) was developed by School of Biology Professor John McDonald and Ph.D. Bioinformatics candidate Vinay Mittal. Details of the pipeline are published in the journal Nucleic Acids Research.

A major bottleneck in the realization of the dream of personalized medicine is no longer technological. It’s computational,” said McDonald, director of Georgia Tech’s newly created Integrated Cancer Research Center. “R-SAP follows a hierarchical decision-making procedure to accurately characterize various classes of gene transcripts in cancer samples.”

There are at least 23,000 pieces of RNA in the human genome that encode the sequence of proteins. Millions of other pieces help regulate the production of proteins. R-SAP is able to quickly determine every gene’s level of RNA expression and provide information about splice variants, biomarkers and chimeric RNAs. Biologists and clinicians will be able to more readily use this data to compare the RNA profiles or “transcriptomes” of normal cells with those of individual cancers and thereby be in a better position to develop optimized personal therapies.

“R-SAP can make 100 million reads in just 90 minutes. Running the program simultaneously on multiple CPUs can further decrease that time.”

R-SAP is open source software, freely accessible at the McDonald Lab website.

“This is another example of Georgia Tech’s ability to merge computer technology with science to create an essential feature of next-generation bioinformatics tools,” said McDonald. “We hope that R-SAP will be a useful and user-friendly instrument for scientists and clinicians in the field of cancer biology.”

Georgia Tech has created a new data analysis algorithm that quickly transforms complex RNA sequence data into usable content for biologists and clinicians.
Algorithms to Search for Previously Undetected Earthquakes in Continuous Seismic Recordings

Research to develop a next-generation computer code for seismic data analysis is being conducted at Georgia Tech in an effort to help advance the scientific community’s capabilities in detecting aftershocks that result from major earthquakes. Continuous recording techniques for seismic activity are amassing data in record volume and the project seeks to improve on an existing computational technique effective in finding previously undetected seismic events within the data.

Dr. Zhigang Peng, Earth and Atmospheric Sciences, and Dr. Bo Hong, Electrical and Computer Engineering, the principal investigators on the project, are developing new algorithms for massive scale analysis that involves thousands of template seismic events – events used as a baseline to find other similar seismic activity – and years of continuously recorded data.

The team is using the existing computational technique known as waveform matched filtering and implementing it on massively parallel HPC platforms. The technique involves an analysis of the stored data and cross-referencing the template seismic events to find previously undetected earthquakes.

Waveform matched filtering is computationally intensive, and the analysis of 3,000 template events, for example, recorded at a handful of seismic stations would take approximately 3,000 CPU hours, or 125 days, to detect events for just one day of input data.

*The computational complexity is a major bottleneck that prevents this technique from being applied at a massive scale,* says Dr. Hong. *We plan to take this challenge and are proposing to accelerate the waveform matched filter technique through GPU computing.*

Working together since fall 2011, the researchers have recently developed a GPU code to substantially increase the speed of the waveform matched filter procedure. The code design targeted HPC platforms equipped with GPU clusters and disk arrays, and have been tested using the multiple GPU workstations at Dr. Hong’s lab and the ‘Keeneland’ GPU clusters at Georgia Tech.

The team is currently applying the algorithms to data collected from moderate to large earthquakes that occurred recently in California and Costa Rica. The magnitude 7.6 earthquake that occurred in Costa Rica in September 2012 was especially interesting, because earthquake scientists have expected it to occur for years, and many seismic instruments have been deployed around the epicenter to capture the main shock and its aftershocks. The successful deployment of the Georgia Tech algorithm to study the Costa Rica earthquake has the potential to create a significant impact within the field of seismology and beyond.

*We envision that this code, once made available, will be used by many researchers in the field of seismology for detection of new seismic events at times and in regions that are otherwise impossible for human analysts to handpick all the earthquakes,* says Dr. Peng.

In addition, the method could be incorporated by seismic data centers to develop automatic earthquake location techniques in near-real time.

Understanding Genomic Evolution with Petascale Computational Tools

Technological advances in high-throughput DNA sequencing have opened up the possibility of determining how living things are related by analyzing how their genes have been rearranged on chromosomes. However, inferring such evolutionary relationships in this way is computationally intensive, even on the most advanced computing systems.

*Genome sequences are now available for many organisms, but making biological sense of the genomic data requires high-performance computing methods and an evolutionary perspective, whether you are trying to understand how genes of new functions arise, why genes are organized as they are in chromosomes, or why these arrangements are subject to change,* said David Bader.

Even on today’s fastest parallel computers, it could take centuries to analyze genome rearrangements for large, complex organisms. Thus, the research team is focusing on future generations of petascale machines, which will be able to process more than a thousand trillion calculations per second, compared to a few hundred thousand per second on the average personal computer.

The researchers, led by Bader, plan to develop new algorithms in an open-source software framework, using parallel, petascale computing platforms to infer ancestral rearrangement events. On a data set of a dozen bellflower genomes, the software determined the flowers’ evolutionary relatedness one billion times faster than the original implementation.

The next test will analyze a collection of fruit fly genomes, providing a relatively simple system to understand the mechanisms that underlie gene order diversity, which can later be extended to more complex mammalian genomes, such as primates.

The researchers believe these new algorithms will make genome rearrangement analysis more reliable and efficient, while potentially revealing new evolutionary patterns. In addition, the algorithms will enable a better understanding of the mechanisms and rate of gene rearrangements in genomes, and the importance of the rearrangements in shaping the organization of genes within the genome.

*Ultimately this information can be used to identify microorganisms, develop better vaccines, and help researchers better understand the dynamics of microbial communities and biochemical pathways,* added Bader.

“...To identify microorganisms, develop better vaccines, and help researchers better understand the dynamics of microbial communities and biochemical pathways.”

- David Bader

Zhigang Peng
Associate Professor, School of Earth and Atmospheric Sciences

Bo Hong
Assistant Professor, School of Electrical and Computer Engineering

David Bader
Professor, School of Computational Science and Engineering; Executive Director for High Performance Computing

Understanding Genomic Evolution with Petascale Computational Tools
Navigating a Sea of Data

As recently as a decade ago, the challenge in data analytics was in gathering adequate amounts of data. Today the challenge lies in making sense of the oceans of data that are now available. Haesun Park is tackling that challenge in her role as the principal investigator of the Georgia Tech team leading the Foundations of Data and Visual Analytics (FODAVA) program, sponsored by the National Science Foundation and the U.S. Department of Homeland Security. In its role as the lead institution of the FODAVA program, Georgia Tech is working with nineteen other universities to develop the data and visual analytics field.

Data is everywhere. Some of the largest data sets these days are used in areas such as network security, health care, bioinformatics, and homeland security. The FODAVA program aims to capitalize on knowledge and expertise in mathematics, computational science, information visualization, and cognitive science to produce new methods to “detect the expected and discover the unexpected in massive data sets.”

The problem is complicated further by the fact that some data doesn’t lend itself to analysis using numerical methods, Park noted. In many fields, tens of thousands of free-form, unstructured documents—such as e-mail exchanges and doctors’ notes—are being collected, and people need a way to systematically analyze them. She is hoping to help find the way.

In many fields, tens of thousands of free-form, unstructured documents—such as e-mail exchanges and doctors’ notes—are being collected, and people need a way to systematically analyze them. She is hoping to help find the way.

“The kind of work I did as a pure mathematician felt too theoretical, too isolated from real life,” said Park. “The work I do now provides very foundational understanding of problems. But at the same time, I try to work very closely with the people in the application domain. Theory provides a foundation for applications, and applications provide important insights into theoretical work.”

“I have worked on some problems where everybody thinks the ultimate solution has already been found,” said Park. “But sometimes I am able to go back and approach it differently and find an even better way to do it. It’s really nice when that happens.”

Design of Next Generation Nuclear Reactor Tied to New Advanced Computational Tool

Woodruff School of Mechanical Engineering and Nuclear and Radiological Engineering researchers are developing a high performance computational analysis tool for the Generation IV Nuclear Energy Systems Initiative (Gen IV) from the U.S. Department of Energy.

Advanced computational tools are a necessary component for developing new reactor systems as part of the DOE’s strategic focus in maintaining the nation’s leading position in nuclear energy research and development. The Institute for Data and High Performance Computing at Georgia Tech has awarded a seed grant to the research team to support their work on the federal initiative.

Srinivas Garimella and Farzad Rahnema, professors in the George W. Woodruff School of Mechanical Engineering, along with Ph.D. student Alex Hunning, are developing a multiprocessor thermal hydraulic analysis tool for the design of the reactor, called the very high temperature reactor (VHTR).

The VHTR is the leading candidate for the Next Generation Nuclear Plant project under the Gen IV initiative and is being designed to meet a federally mandated launch in 2021.

Analysis of the heat transfer, temperature distributions, and coolant flow through the reactor core is required for both operational design enhancement and transient safety qualification. The only way to study these processes is through computations that can lead to improved VHTR designs. The advantage of the Georgia Tech team’s computational methodology over existing techniques is an improved level of 3-D modeling detail for the whole core.

“The primary computational challenge associated with determining the core temperature and coolant distributions is due to the large physical or problem scale,” says Rahnema. “Since temperatures are expected to widely vary across the core, it is important to predict the temperature within each fuel pin to ensure temperature limits are not exceeded.”

To realistically model the temperature of each fuel pin and coolant channel, it requires a million or more unique computational cells that divide the physical geometry and materials into smaller pieces.

“The single complex task now becomes a large set of many simpler tasks,” says Garimella. “This set of computations is prohibitively large for traditional single CPU desktop machines.”

Parallelization and implementation of the thermal hydraulic scheme across many CPUs will work to overcome this obstacle. In addition to multiple CPUs, GPU computing techniques are being employed to take advantage of specific or attractive hardware differences resulting in a state-of-the-art, fast running, thermal hydraulic analysis tool for the VHTR.

The project’s first phase involves developing a parallelized thermal hydraulic tool for VHTR analysis. The second phase couples additional physical models to capture nuclear-thermal feedback effects. Both phases require a multidisciplinary effort from high performance computing, traditional nuclear sciences and nuclear design, thermal fluid sciences and thermodynamics.

The parallel computer code is being deployed on the CUDA architecture, developed by NVIDIA, for parallel GPU computing processes. High performance computing will be accomplished through the use of the PACE advanced computing environment at Georgia Tech using a portion of the FoRCE cluster that consists of 9 compute nodes with 240 total CPUs and 480 GPU cores per node.
Keeneland

Several recent studies have identified the development of methods for efficiently programming many-core architectures as a major challenge. The Keeneland Project has developed an experimental platform using many-core processors with very high memory bandwidth and graphics processing units for use as an experimental high-performance parallel computing.

Led by Jeffrey Vetter, joint professor of the School of Computational Science and Engineering at Georgia Tech and Oak Ridge National Laboratory, Keeneland is supported by the National Science Foundation and is operated in partnership with the University of Tennessee-Knoxville and Oak Ridge National Laboratory.

Keeneland initially deployed a small, experimental, high performance computing system consisting of a system from Hewlett-Packard with attached NVIDIA accelerators in 2010.

The project upgraded the heterogeneous system to a larger and more powerful system based on a next-generation platform and NVIDIA accelerators. The Keeneland Full Scale system (KFS) (792 GPUs, ~515 TFLOPS) successfully completed extensive acceptance testing in fall 2012. This Georgia Tech-led project is the largest GPU-based supercomputer resource that the National Science Foundation owns.

Keeneland will be integrated into the TeraGrid to make the system more widely accessible to the research community. The project team will use this system to develop scientific libraries and programming tools to facilitate the development of science and engineering research applications. The project team will also provide consulting support to researchers who wish to develop applications for the system using OpenCL, or to port applications to the system. Currently, the system has over 125 projects from around the United States running on it. The KFS will be the Extreme Science and Discovery Environment’s only GPU capacity computing resource in the near future.

CUDA Center of Excellence

In 2010, NVIDIA designated Georgia Tech as a CUDA Center of Excellence. Georgia Tech is among ten other universities and research organizations in the United States and abroad that were given this distinction.

Jeffrey Vetter serves as principal investigator of the CUDA Center of Excellence. “Georgia Tech has a long history of education and research that depends heavily on the parallel processing capabilities that NVIDIA has introduced with its CUDA architecture,” Vetter said. “This award allows us to focus, what is now a large amount of activity across twenty-five different research groups, under a single center, which will significantly amplify our research capabilities.”

CUDA is NVIDIA’s parallel computing architecture that enables dramatic increases in computing performance by harnessing the power of a GPU. Georgia Tech is engaged in a number of research, development and educational activities that leverage GPU computing. These activities span the full gamut: applications, software development tools, system software, and architectures.

GTRI Big Data, Analytics and High Performance Computing (HPC) Strategic Initiative

The Georgia Tech Research Institute (GTRI) is helping to define the direction and path of massive data research at Georgia Tech through the development of the Big Data, Analytics and High Performance Computing (HPC) Strategic Initiative.

GTRI is working to develop an institute-level Big Data group that uses its knowledge base to “define what Big Data problems look like, work with applied research projects to identify needs for solutions and in turn help the group become smarter in solving problems in the space,” says Margaret Loper, principal researcher scientist in GTRI, who is leading the initiative along with GTRI principal research scientist Dan Campbell. GTRI has also brought in experts from six of its labs to help establish a set of operational definitions for the research space and identify the technologies that could set Georgia Tech apart to take a lead in Big Data research.

“Collection of data is the trend of where everything is going, and organizations should be strategically focused to understand the growth in data and the computational need to process that data,” says Loper.

To develop the collaborative framework, expertise through other research centers - such as the Institute for Data and HPC - is being leveraged. GTRI has also initiated Big Data “requirements analyses” for research projects. In several of GTRI labs, its Cyber Lab, researchers are developing a malware repository and the subject-matter experts in cyber-related problems are collaborating with researchers in the Big Data group to understand the innovation needed to do analytics with a very quick response time. GTRI is also collaborating with College of Computing faculty in another strategic area - computational social science - where the rise of online social networks has made the research field a good fit for Big Data.

Digital Pathology Collaboration with Emory Uses Fast, GPU-enabled Machine Learning for Medical Data

Georgia Tech and Emory University researchers are collaborating to help advance next-generation medicine using machine learning methods. A large-scale digital pathology project, one of three initial projects, is part of a broader collaborative effort between the two institutions established to address specific computational challenges in medical science. The research seeks to create greater automation in the analysis of high-resolution medical images and related clinical data used by medical practitioners that will impact both costs and diagnostic accuracy.

“We will enable this by making morphological classification possible on massive image datasets,” said Joel Saltz, the project’s principal investigator, and chair and professor of Emory’s Department of Biomedical Informatics.

Alexander Gray, the Georgia Tech co-PI, said that the team will address the design, implementation, and application of new distributed memory and multi-core CPU-GPU algorithms for machine learning methods, specifically for clustering. The project will study a fundamental open problem in scalable machine learning, and the resulting tool will be used to obtain the largest such computation in the medical image analysis literature.

To accelerate the machine learning method of clustering, “k-means” will be used as the base algorithm, which is a mainstay approach to finding clusters in data. This will be applied to automatically finding groups corresponding to disease subtypes and states using micro-anatomic features, such as nuclei and cells, extracted from high resolution pathology images. Software is expected to be developed in a reusable fashion as part of MLPACK - an open source scalable machine learning library developed at Georgia Tech - and the new joint GT/Emory open source software product will seamlessly interoperate with MLPACK. This new software will be called GPU-MLPACK and will house GPU code, which can be seamlessly called from MLPACK.

The research team includes Georgia Tech members Alexander Gray, Bill March, James Cline, Dongyeeol Lee and Ajinkya Kale and Emory members Joel Saltz, Tahsin Kurc, Michael Nalisnik, Tony Pan, George Teodosio, Lee Cooper, and Jun Kong.
Invited Keynote and Plenary Talks at Major Conferences

Haesun Park (CSE), gave the plenary keynote lecture, “Nonnegative Matrix Factorization: Algorithms and Applications,” at the SIAM International Conference on Data Mining (SDM11), Mesa, Arizona, April 28-30, 2011.


C.F. Jeff Wu (IE) was invited to give the prestigious American Statistical Association’s 2012 Deming Lecture: “Quality Improvement: From Autos and Chips to Nano and Bio.” San Diego, CA, July 31, 2012.


Steve Harvey (Biology) gave the Plenary Lecture “Satellite Tobacco Mosaic Virus: An All-Atom Structure ... and then a Surprise!” at the Workshop on Physical Virology at the Abdus Salam International Centre for Theoretical Physics, Trieste, Italy, Sept. 24-28, 2012.

Major Grants and Funding Awards

DARPA, “Proactive Detection of Insider Threats with Graph Analysis at Multiple Scales,” PIs: T. Senator (SAIC) and D.A. Bader (GTRI), Anomaly Detection at Multiple Scales (ADAMS) Program, 5/1/2011 – 4/30/2013.

DARPA, "SOCINet: Social Signals for Identifying & Using Subnetworks," PIs: E. Gilbert (IC), co-Pi: D.A. Bader (CSE), I. Essa (IC), C. Devorel (CS), L. Weiss (GTRI), and E. Bristoe (GTRI), Social Media in Strategic Communication (SMiSC) Program, 2/1/2012 – 1/31/2013.


DARPA, “Fast Algorithms on Perfectly Heterogeneous Distributed Data for Interactive Analysis,” PIs: P. Fujimoto (CSE), co-Pis: Richard Boyd (GTRI), Barry Drake (GTRI), Alexander Gray (CSE), Guy Lebanon (CSE), Haesun Park (CSE), Hongyuan Zha (CSE), XDADA Program, 2012-2017.


NSF 1148310, "A Glass Box Approach to Enabling Open, Deep Interactions in the HPC Toolchain," PI: Karsten Schwan (CS), Co-PIs: Sudhakar Yalamanchili (ECE), Greg Eisenhauer (CS), Matthew Wolf (CS), Software Institutes Program, 6/1/2012 – 05/31/2015.


Major Research Accomplishments


Jason Riedy (CSE) developed scalable code for detecting communities in graphs and social networks that won the 10th DMACCS Implementation Challenge’s graph clustering mixed-criteria quality and Pareto challenges; the code processed a real-world graph with 3.3 *billion*edges in less than 10 minutes.

Joshua S. Weitz (Biology & Physics) and collaborators at the University of Illinois Urbana-Champaign have developed a multi-scale model of host-viral evolution that integrates molecular, ecological, and evolutionary parameters. Working with the National Center for Supercomputing Applications, the team will port the code to high-performance computing resources and perform large-scale simulations of the co-evolution of hosts and pathogens. The project builds upon work published in Evolution in July 2012.

Georgia Tech’s Center for Experimental Research in Computer Systems (CERCIS), led by Karsten Schwan (CS), has successfully integrated its FlexIO middleware into the ADIOS I/O infrastructure widely used on petascale machines, giving Georgia Tech’s open source software full integration into the ADIOS family.

Honors, Awards and Other Recognition

Best Paper Award: “Improving Prediction Accuracy of Protein-DNA Docking with GPU Computing,” Bo Hong (ECE), Jadong Wu (ECE), and Jun-tao Guo (UNC Charlotte), 2011 IEEE International Conference on Bioinformatics and Biomedicine (BIBM11), November 2011.


David A. Bader (CSE) serving as chair, SIAM Activity Group on Supercomputing (SIAM/SC), January 2012 - December 2013.

David A. Bader (CSE) serving as chair, IEEE-CS Awards Committee, Institute of Electrical and Electronics Engineers (IEEE) Computer Society (2012).

Aparna Chandramowlishwaran (CSE PhD) is a recipient of a 2012-2013 Intel PhD Fellowship.

Alex Gray (CSE) co-founded Skytree Inc., which provides the first general purpose machine learning and advanced analytics engine, designed to accurately process massive datasets at high speeds.

Richard Vuduc (CSE) is the program committee co-chair (joint with Saman Amarasinghe of MIT) for the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP) for 2013.

Richard Vuduc (CSE) is the program committee co-chair (joint with Michael Henoux of Sandia National Laboratories) for the Applications Track of the ACM/IEEE International Conference for High-Performance Computing, Networking, Storage, and Analysis (SC) 2013.