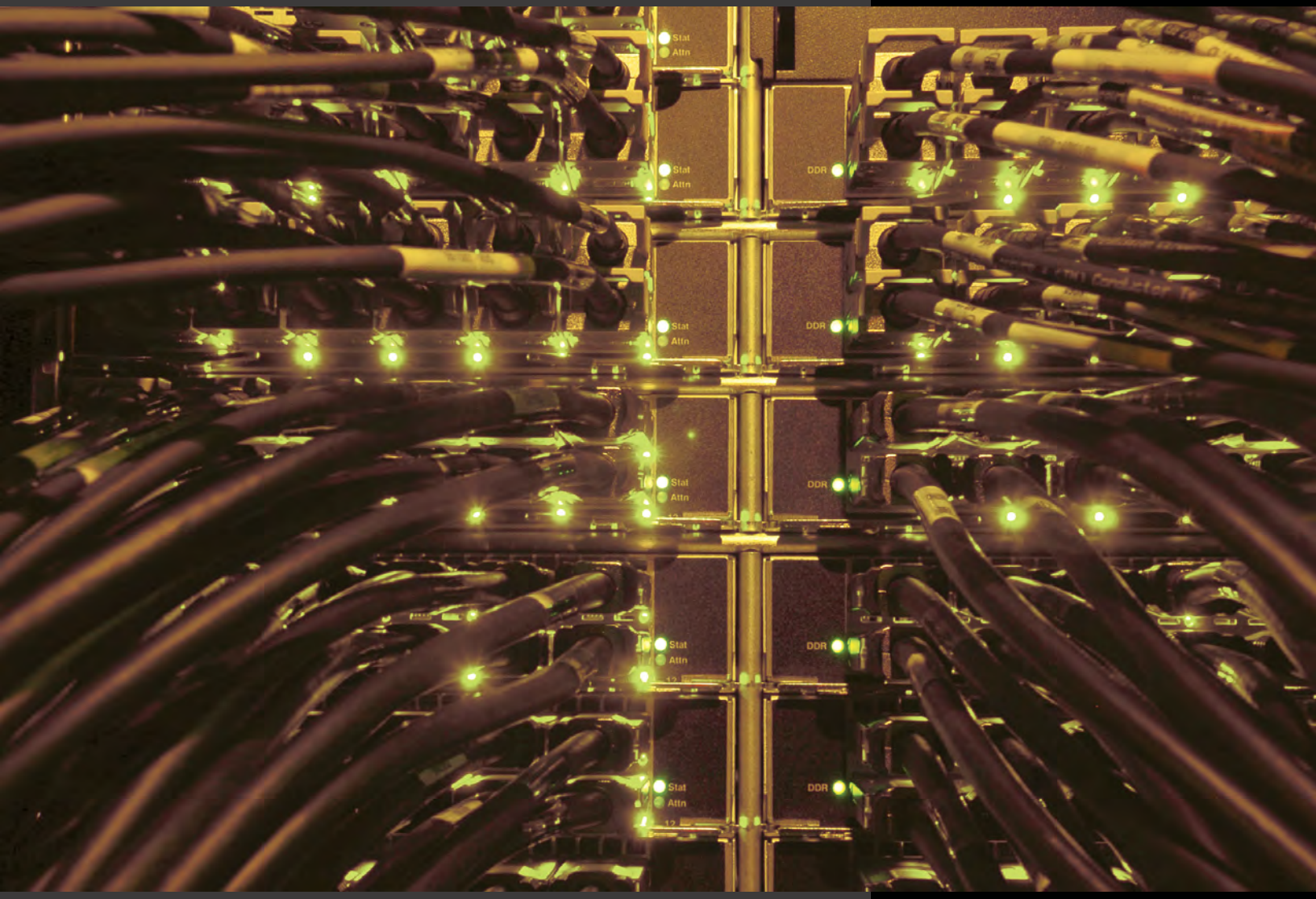
Georgia Tech

Institute for Data and High Performance Computing



Innovation Starts Here

Growing the Research Enterprise

Big Data and High Performance Computing (HPC) are highly interdisciplinary fields that are evolving quickly and having fundamental impact in our technology-driven society. The Institute for Data and HPC at Georgia Tech is developing and leveraging research in these fields to steer the course of innovation and create new capabilities in computing advancement and economic growth.

IDH is creating the technologies that are propelling Big Data and HPC forward. Through its NextGen Codes Initiative IDH has seeded the development of: scalable machine learning libraries, quantum chemistry codes, seismic-detection algorithms, and GPU-enabled computing tools for nuclear reactor designs, among others. IDH launched two new centers in the past year: the Center for Data Analytics and the Center for High Performance Computing: From Big Data to Exascale Computing. These activities are creating synergies between computational experts and domain specialists and help to define Georgia Tech's competitive advantage.

Community Building and Partnerships

IDH is helping to define a future path for massive data and HPC research at Georgia Tech by developing networks of researchers and industry partners and fostering new opportunities in these burgeoning fields. As a partner in the Georgia Tech Research Institute's Big Data, Analytics and HPC Strategic Initiative, IDH is helping foster connections with academic researchers across campus in areas such as sustainability and energy, materials and manufacturing, bioscience and biomedicine, and security. IDH is actively exploring how to integrate Big Data and HPC into the larger Georgia Tech research culture to become a catalyst for innovation. Collaborations with industry partners and national sponsors are a key part of growing the research enterprise. Members of the research community have partnered with several federal agencies and Fortune 500 companies to meet their needs in a number of areas.

Technical Support Across Domains

IDH is taking on a specialized role to make computational advances readily available to domain scientists through software libraries and tools. This allows researchers to substantially invest in these tools, transfer them to new research pursuits, and gain access to a rich repository of managed Big Data and HPC technology. IDH supports the Keeneland Full Scale System, a Georgia Tech-led project establishing the most powerful GPU supercomputer available for research through NSF's Extreme Science and Engineering Discovery Environment (XSEDE) program.

Events, Training and Education

IDH supports collaboration and innovation by bringing together practitioners and thought leaders from Georgia Tech's leading areas through research events and activities. IDH provides support for training that advances researchers' opportunities to apply new technical skills and best practices. The institute offers short courses, workshops and distance education opportunities with industry-leading groups and research organizations. IDH enables knowledge transfer into targeted domain areas and creates impact by helping researchers identify real-world problems and providing solutions with state-of-the-art methodologies and tools.

Richard Fujimoto:
IDH Interim Director

Institute for Data and High Performance Computing

Positioning for Future Research Opportunities

IDH is Georgia Tech's external face for massive data and HPC activities and research on campus. A dynamic research community is visible to a variety of stakeholders through IDH, enhancing Georgia Tech's ability to collaborate with industry and government. IDH demonstrates Georgia Tech's commitment to Big Data, one of 12 core research areas shaping Tech's vision to be a leader in the regional innovation ecosystem. IDH supports industry events including a tradeshow booth at the annual Supercomputing Conference to enhance Georgia Tech's visibility and marketing efforts in Big Data and HPC. IDH also develops and maintains news and marketing initiatives related to Georgia Tech projects and researchers advancing the Big Data and HPC fields.

Contact:

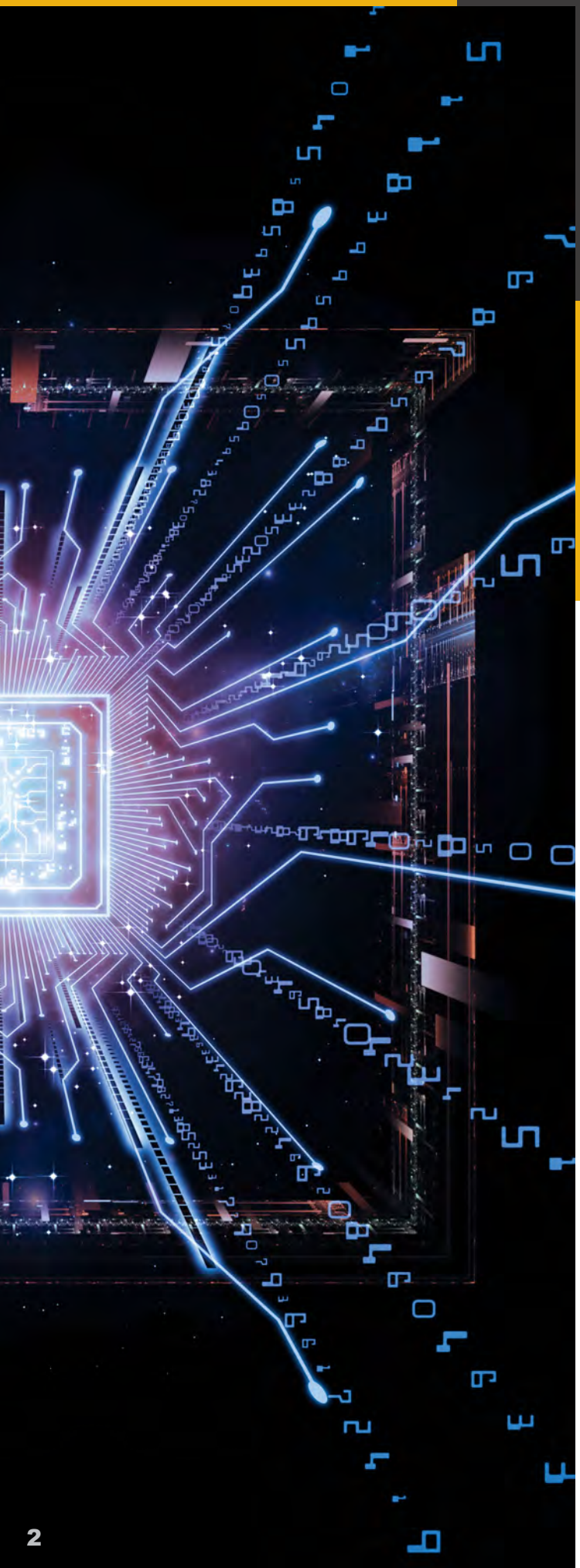
IDH Interim Director
Richard Fujimoto, Ph.D.
Computational Science & Engineering
fujimoto@cc.gatech.edu

Institute for Data and High Performance Computing
Georgia Tech
Klaus Advanced Computing Building
266 Ferst Drive NW
Atlanta, GA 30332-0280

404.385.4785
Fax: 404.385.7337
idh.gatech.edu

"IDH provides a platform for innovation in Big Data and HPC to tackle society's most important and challenging problems."

TABLE OF CONTENTS



XDATA Aims to Extract Knowledge from Growing Digital Data

Georgia Tech Team Wins $2.7 Million Award to Advance Big-Data Technology for DARPA

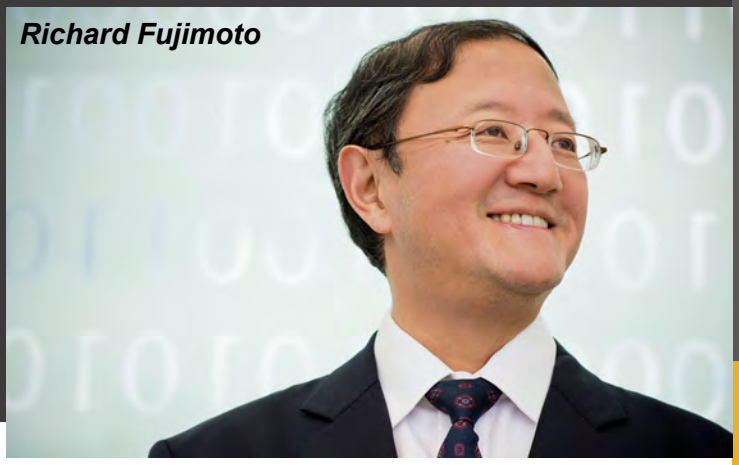
The XDATA award is part of a $200 million multi-agency federal initiative for Big-Data research and development announced in March 2012. The initiative is aimed at improving the ability to extract knowledge and insights from the nation's fast-growing volumes of digital data. Numerous Big-Data-related research endeavors are underway at Georgia Tech.

A research team led by **Haesun Park**, Professor in the School of Computational Science and Engineering and Director of the newly formed Center for Data Analytics, received a $2.7 million award from the Defense Advanced Research Projects Agency (DARPA) to develop technology intended to help address the challenges of Big Data. The contract is part of DARPA's XDATA program, a four-and-half year research effort to develop new computational techniques and open-source software tools for processing and analyzing data, motivated by defense needs.

Selected by DARPA to perform research in the area of scalable analytics and data-processing technology, the team focuses on producing novel machine-learning approaches capable of analyzing very large-scale data. In addition, team members are pursuing the development of distributed computing methods that can implement data-analytics algorithms very rapidly by simultaneously utilizing a variety of parallel-processing environments and networked distributed computing systems.

The algorithms, tools and other technologies developed will be open source, to allow for customization. Under the open-source paradigm, collaborating developers create and maintain software and associated tools. Program source code is made widely available and can be improved by a community of developers and modified to address changing needs.

The Georgia Tech XDATA effort builds upon foundational methods and software developed under the Foundations of Data and Visual Analytics (FODAVA) research initiative, a 17-university program led by Georgia Tech and funded by the National Science Foundation and the Department of Homeland Security.



Richard Fujimoto

The FODAVA effort has produced multiple innovative visual analytics software systems such as the FODAVA research test bed, Visual Information Retrieval and Recommendation System (VisIRR) and other tools for interactive classification, clustering and topic modeling tasks.

"The FODAVA document retrieval and recommendation system uses automated algorithms to give users a range of subject-search choices and information visualization capabilities in an integrated way, so that users can interact with the data and information throughout the problem-solving process to produce more meaningful solutions," said Park, who is also FODAVA's director. "For XDATA, we will enhance these visualization and interaction capabilities and develop distributed algorithms that allow users to solve problems faster and on a larger scale than ever before."

Also participating from the School of Computational Science and Engineering are Professor **Hongyuan Zha** and Research Scientist **Jaegul Choo**, who has previously led development of visual analytics systems on the FODAVA project. Investigators from the Georgia Tech Research Institute (GTRI) also contribute to the XDATA initiative. Senior Research Scientists **Barry Drake** and **Richard Boyd** are responsible for handling the computational demands of implementing the data analytics algorithms being developed.

GTRI's task involves enabling these algorithms to run on a networked distributed computing system. By configuring the software to operate on multiple processors simultaneously, the researchers believe they can ensure that the algorithms solve large-scale problems very rapidly—a requirement of the DARPA award. The GTRI team is applying the latest advances in high performance numerical libraries to speed up the underlying computations of the higher-level data analytics algorithms and building tools to integrate the data analytics into a professional open-source package.

FLAMEL: From Learning, Analytics, and Materials to Entrepreneurship and Leadership Doctoral Traineeship Program

Georgia Tech to Exploit Big Data for Accelerating Materials Design and Manufacture through FLAMEL Traineeship Program

Georgia Tech has been awarded $2.8 million from the National Science Foundation to start a program to train a new type of data scientist capable of creating advanced materials and bringing them to market at a fraction of the time it now takes, typically 15 to 20 years.

"The goal of this program is to employ advances in 'big data' and information technology to significantly reduce the timelines now required for new materials to be created and incorporated into commercial products," said School of Computational Science and Engineering Chair and Regents' Professor **Richard Fujimoto**, the principal investigator for the grant.

"The program will be transformational in bringing 'big data' researchers together with materials scientists, engineers, and mathematicians to quantify the microstructures that comprise materials and develop new algorithms and software for their design," said Fujimoto.

The five-year program will provide funding for 24 doctoral trainees but is expected to create educational opportunities that will impact hundreds of Georgia Tech students.

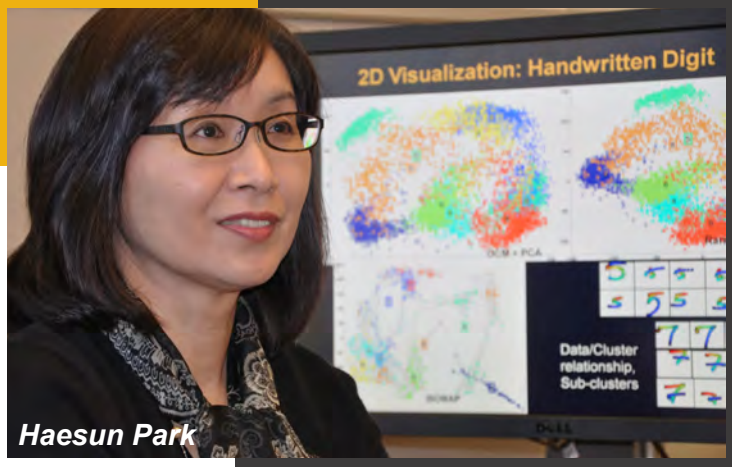
The new program includes a focus on entrepreneurship to enable graduate trainees to transform technical innovations into commercial products and services. Called FLAMEL—From Learning, Analytics, and Materials to Entrepreneurship and Leadership, the program will leverage Georgia Tech's recent investment in MatIN, a cyberinfrastructure platform designed to enable rapid interdisciplinary collaboration in materials development and manufacture. More information is available at **flamel.gatech.edu**.

The program is funded through NSF's Graduate Education and Research Traineeship (IGERT) program, award number DGE-1258425.

New Research Centers are Foundation for Innovation

Two new research centers spawned by IDH are becoming leaders among Georgia Tech's emerging technology groups at the heart of research innovation in Big Data and HPC.

The Center for Data Analytics (CDA) is Georgia Tech's newest lead in the design and development of a unified community for Big Data and analytics. The center, led by Professor **Haesun Park**, School of Computational Science and Engineering, is focused on enabling scientific advancement of today's challenging Big Data problems and providing integrated leadership on data analytics. CDA is harnessing the institute's strengths by bringing together a large number of faculty practicing foundational disciplines in the Big Data research space.



Haesun Park

A broad and deep range of subject-matter expertise—including machine learning, modeling and simulation and data visualization—will allow collaborators regionally and globally to investigate new directions anywhere within scientific inquiry. CDA is designed to bring more opportunities to Georgia Tech from external scientists, government agencies and large corporations to tackle scientific problems that cannot be solved without emerging computational toolsets.

FODAVA Defines Foundations for Visual Analytics Research Field

The Foundations on Data Analysis and Visual Analytics (FODAVA) research initiative, a 5-year project led by Georgia Tech and under the direction of Professor **Haesun Park**, School of Computational Science and Engineering, is the national genesis for defining the computational foundations for data and visual analytics fields. FODAVA is engaged in a collaborative effort among 17 partner institutions funded jointly by the National Science Foundation and the Department of Homeland Security.

FODAVA performs foundational research and investigates ways to improve visual analytics of massive data sets through advances in areas such as machine learning, numeric and geometric computing, optimization, computational statistics and information visualization.

The initiative established data and visual analytics as a distinct research field and built a dynamic community of researchers collaborating through foundational research, research workshops, conferences, industry engagement and technology transfer. FODAVA-generated collaborations have continued to build, and technical reports as well as other research data from the national initiative may be found at **fodava.gatech.edu**.

New Tools Weight Connections Between Large-Scale Data in High Dimensional Space

A research team led by Computational Science and Engineering Professor **Haesun Park** is taking on the Big Data challenge through an approach that finds unseen structures in vast amounts of high dimensional data and represents them in the limited two-dimensional screen space, allowing interaction with the data to achieve more meaningful solutions. This new foundational method for clustering data is based on characterization of the space in which the data resides and simultaneously reduces the dimensions, so that the data is represented using only a smaller number of collapsed and informative features.

One example might be a database of research paper collections too numerous for manual viewing. Using a tool such as Georgia Tech's UTOPIAN (User-driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization), the documents are simultaneously categorized into clusters (e.g., articles about similar topics stay together) while incorporating the user's prior knowledge through visual interaction. Park believes this clustering method could be applied in a number of ways because it captures meaningful cluster structure more naturally than completely automated methods by allowing human knowledge to guide the solution process.

The Center for High Performance Computing (HPC): From Big Data to Exascale Computing, led by Executive Director **David A. Bader**, is designed to bring together Georgia Tech's interdisciplinary research and education activities in high performance computing in order to better leverage Georgia Tech's capabilities in this area and enable solving grand challenges in computational science and engineering.


David A. Bader

The center is strategically focused to meet the demands of federal priorities from multiple agencies over the next five years that are focused on high performance computing and Big Data. The center recognizes that the expanding cyberinfrastructures within organizations provide for new HPC capabilities, which are essential, not optional, to the aspirations of research communities.

The Georgia Tech Center for HPC is focused on being a dominant innovator in HPC and massive data technology and a creator of software and tools enabling and accelerating discovery and innovation in targeted application domains.

DARPA Awards Georgia Tech Energy-Efficient High-Performance Computing Contract

Georgia Tech is in the first phase of a cooperative agreement contract from the U.S. Defense Advanced Research Projects Agency (DARPA) to create the algorithmic framework for supercomputing systems that require much less energy than traditional high-speed machines, enabling devices in the field to perform calculations that currently require room-sized supercomputers.

Awarded under DARPA's Power Efficiency Revolution for Embedded Computing Technologies (PERFECT) program for $561,130—for phase one of a negotiated three-phase $2.9 million contract—the cooperative agreement contract is one piece of a national effort to increase the computational power efficiency of "embedded systems" by 75-fold over the best current computing performance in areas extending beyond traditional scientific computing. Computational Science and Engineering Professor **David Bader** is principal investigator on the Georgia Tech cooperative agreement, along with research scientist and co-PI **Jason Riedy**. The project bears the acronym GRATEFUL: "Graph Analysis Tackling power-Efficiency, Uncertainty and Locality."

Such a system would have benefits in energy conservation and improve tactical advantages of supercomputing in military situations.

New Georgia Tech Software Recognizes Key Influencers Faster Than Ever

Determining the most influential person in a social media network is complex. Thousands of users are interacting about a single subject at the same time and new people are constantly joining the streaming conversation.

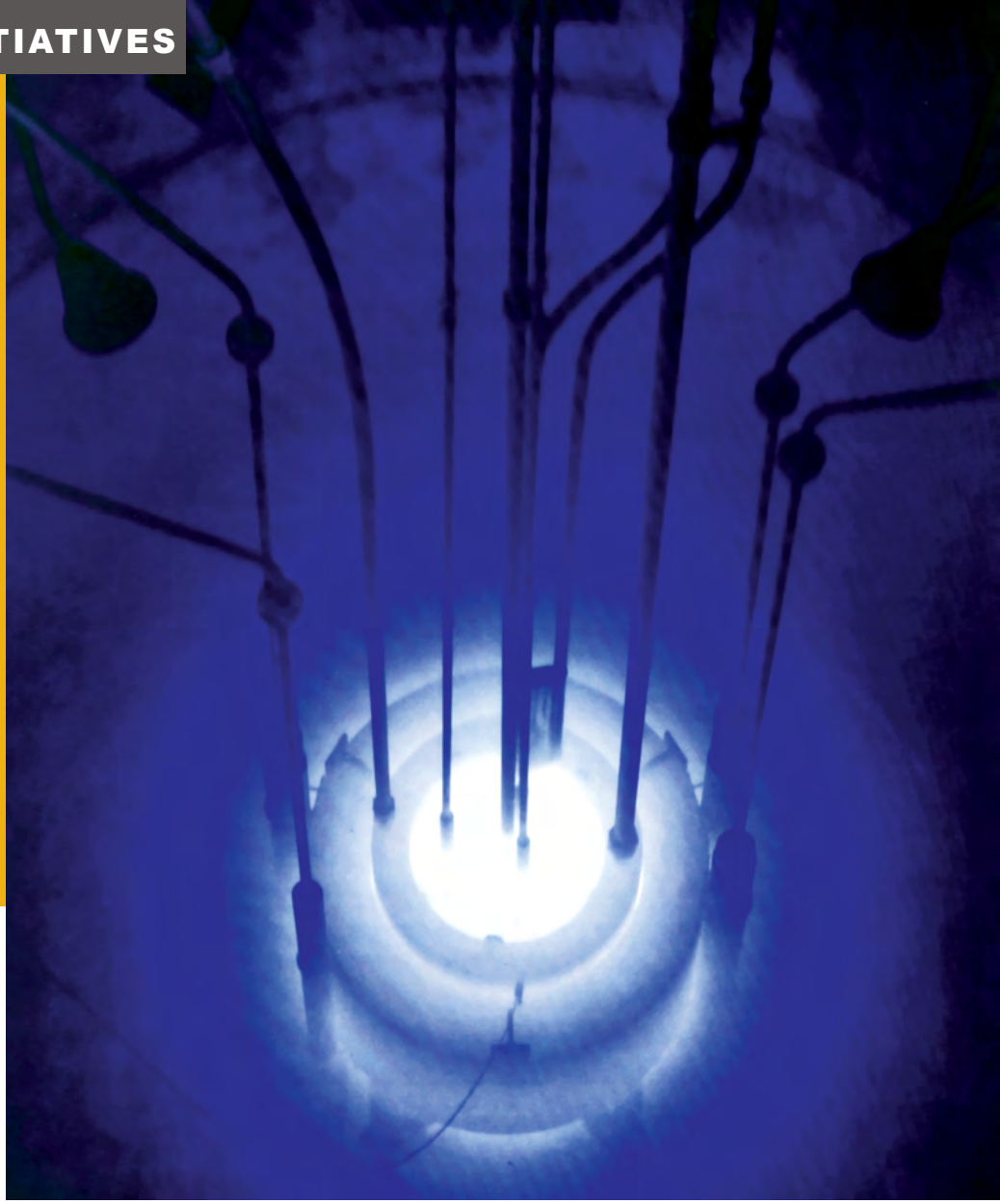
A Georgia Tech team, led by Computational Science and Engineering Professor **David Bader**, has developed a new algorithm that quickly determines the most connected person in a network. The algorithm can identify influencers as information changes within a network, making it a first-of-its-kind tool. The algorithm takes existing graph (network) data and does the bare minimal computations affected by the inserted edges or connections.

The measurement for how connected someone may be in a graph can be computed more than 100 times faster in some cases using the Georgia Tech software.

"Despite a fragmented social media landscape, data analysts would be able to use the algorithm to look at each social media network and mark inferences about a single influencer across these different platforms," said Bader.

The project is supported by the National Science Foundation (Award Number CNS-0708307).The open source software is available to businesses.

Interdisciplinary Georgia Tech teams in various research areas, such as chemistry, nuclear energy, earthquake detection and large-scale data analysis, are tackling the biggest challenges in society through computational problem-solving. Georgia Tech is focused on developing the next generation of "big data" software and high-performance computing architecture that advances applications across the globe.



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 with seed funding from IDH for the Generation IV Nuclear Energy Systems Initiative (Gen IV) from the U.S. Department of Energy.

Professor **Srinivas Garimella** and Professor **Farzad Rahnema**, associate chair of the Woodruff School and chair of the Nuclear and Radiological Engineering and Medical Physics Program, along with Ph.D. student **Alex Huning**, have completed the first phase of the project to develop a multiprocessor thermal hydraulic analysis tool for the design of the reactor, called the very high temperature reactor (VHTR).

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 Georgia Tech methodology employs an improved level of 3-D modeling detail for the whole core. 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.

Parallelization and implementation of the thermal hydraulic scheme across many CPUs will work to meet the computational requirements of the project. GPU computing techniques take advantage of specific or attractive hardware differences, resulting in a state-of-the-art, fast running, thermal hydraulic analysis tool for the VHTR.

How Much Time, Energy and Power Will Your Computation Need?


Richard Vuduc

Computational Science and Engineering Associate Professor **Richard Vuduc** is exploring ways to design algorithms and software that are not only fast, but also power- and energy-efficient. This research, conducted in his lab, the HPC Garage, aims to help supercomputers and data centers run "greener" and also make mobile batteries last longer.

As a first step, **Jee Whan Choi**, a senior graduate student in the HPC Garage, has developed the energy roofline model, which predicts the best-case time, energy and power required by a computation, given the machine on which it is executed.

"The energy roofline tells an algorithm designer or programmer the best he or she can do on a machine," Vuduc explains.

The model can be used to compare systems as well. For example, many researchers in HPC are asking whether a supercomputer built from low-power mobile graphics co-processors, or GPUs, can outperform one built from high-end desktop or server-class GPUs. Vuduc's team has found the energy roofline predicts the high-end system has a clear power and performance advantage for compute-heavy workloads, but the low-power designs can have an advantage for data-intensive workloads.

Beyond these initial results, Vuduc says the next step is to investigate how the model may be used to automatically tune ("autotune") code to improve energy and power efficiency.

Algorithm for Continuous Seismic Recordings Finds Previously Undetected Earthquakes

Current research to develop a next-generation computer code for seismic data analysis will help advance capabilities in detecting aftershocks that result from major earthquakes.

Associate Professor **Zhigang Peng**, School of Earth and Atmospheric Sciences, and Assistant Professor **Bo Hong**, School of 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 will use an existing computational technique known as waveform matched filtering and implement 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 aftershocks.

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.

"We plan to take this challenge and are proposing to accelerate the waveform matched filter technique through GPU computing," says Hong.

Advanced Chemistry Code Reduces Computation Time With Top Speeds

The most advanced computer program for quantum chemistry to date, PSI4, was recently released and researchers anticipate it to be one of the most popular quantum chemistry codes for simulations of large molecules in the near future.

The goal of the PSI4 project is to develop efficient, parallel algorithms incorporating the very latest numerical approximations in quantum chemistry, including various rank-reduction techniques to reduce computation times and resource requirements.

Principal investigators **David Sherrill**, Professor in the School of Chemistry, and **Edmond Chow**, Associate Professor in the School of Computational Science and Engineering, say the PSI4 team is the first to pursue concerted development of parallel algorithms for the very latest numerical techniques in quantum chemistry.

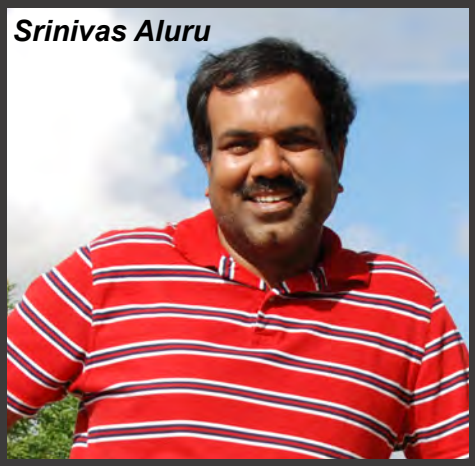
"The [IDH] funding will allow us to make the code scale to a larger number of processors running in parallel, which should allow some of the largest computations performed with accurate, many-body methods," Sherrill noted.

PSI4 was developed collaboratively by Georgia Tech, Virginia Tech, the University of Georgia and Oak Ridge National Lab.

IDH support helped enable the project to receive funding from the National Science Foundation. (Award No. CHE-1011360).

Research projects in the sciences and biomedical fields are a growing part of Georgia Tech's portfolio for "Big Data" innovations. New faculty and new collaborations, including a strategic partnership with Emory University, are advancing these computational solutions to key scientific problems.

Genomes Galore: Big-Data Analytics for High-throughput DNA Sequencing


Srinivas Aluru

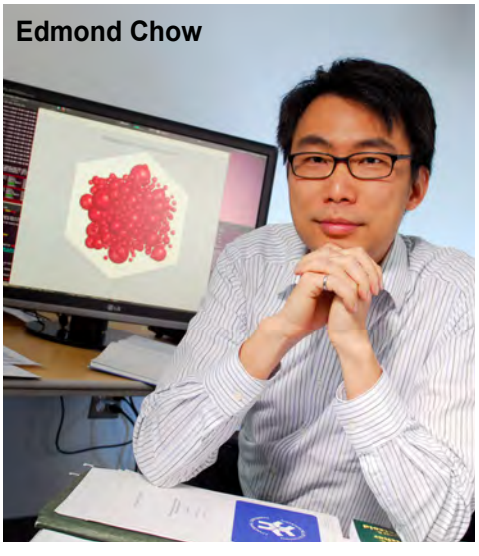
Professor Srinivas Aluru, a newly hired senior faculty member in the School of Computational Science and Engineering, is leading a midscale Big Data project to provide parallel bioinformatics methods for high-throughput sequencing.

Funded with $2 million by the National Science Foundation and the National Institutes of Health as part of the $200 million multi-agency federal initiative for Big-Data research and development, Aluru's work is driven by the need to address grand challenges in human genetics, agricultural biotechnology, and assessment of biological threats, all of which are pursued through sequencing and analysis of genomes.

The Big Data challenge arises from the invention of high-throughput sequencing instruments, which increase throughput by a factor of a billion, while reducing costs by a factor of million. As a result, individual investigators are now able to independently generate a high volume of data, which just a decade ago, could only be achieved by a network of major sequencing centers.

Through this transformational research, Aluru and his team are developing parallel algorithms for various applications of next generation sequencing technologies on a variety of high performance computing platforms. These contributions will be delivered to the broader research community using a multi-layered software stack that can be easily exploited by other researchers and software developers, even by those without HPC expertise.

Large-scale Computer Simulations Show Factors Affecting Molecule Motion in Cells


Edmond Chow

Using large-scale computer simulations, Associate Professor in the School of Computational Science and Engineering **Edmond Chow** and collaborator **Jeffrey Skolnick**, Professor of Biology, are working to simulate how molecules move through the crowded environment inside living cells. Their models incorporate hydrodynamic interactions, believed to be the most important factor in intracellular diffusion.

A detailed understanding of the interactions inside cells—where macromolecules can occupy almost half of the available space—could provide important information to developers of therapeutic drugs and lead to a better understanding of how disease states develop. Ultimately, researchers hope to create a complete simulation of cellular processes to help them understand a range of biological issues.

Sponsored by the National Science Foundation and the National Institutes of Health, this research aims to develop methods for efficiently simulating the motion of macromolecules in an entire cell by employing Stokesian Dynamics

with Brownian motion using advanced scalable algorithms and efficient parallel codes. Using computer simulation provides a reasonably faithful caricature, in which answers to many questions may be found, helping researchers understand what's going on inside a cell. These answers might one day help drug designers better understand how therapeutic compounds work within cells, for instance, or allow cancer researchers to see how cells change from a healthy state to a disease state.

Computationally Quantifying Cell Types in Transplant Patients Provides Insight into Immune System

A Georgia Tech and Emory research team is leveraging its combined expertise to gain unprecedented insights into the dynamics of the immune system following transplants by using newly devised computational techniques. The team is examining how to quantify the cells of various types with sophisticated computational clustering methods and also to infer from the clustered data the causal, dynamic control systems that lead to the unexplained and widely differing trends in different cell types and drastically different patient outcomes.

Following transplants, the composition of T-cell subtypes in a patient's immune system varies as a consequence of the patient's genuine biological features and of different medical treatment regimens. The team has at its disposal a growing body of data consisting of various types of cells that have been and are being quantified at multiple time points following a transplant.

The team is using medical data to develop a better classification of clusters of cell types, improve methods for tracking clusters along time horizons and gain specific insights into classes of longitudinal cluster trends associated with favorable or unfavorable patient outcomes. The research will also develop preliminary dynamic models that represent different population trajectories in a satisfactory manner.

The primary collaborators from each institution are **Eberhard O. Voit**, Professor and David D. Flanagan Chair of the Wallace H. Coulter Department of Biomedical Engineering at Georgia Tech and Emory University; Ashish Sharma, Assistant Professor in Biomedical Informatics at Emory University; and Sharath Cholleti, Research Scientist in the Center for Comprehensive Informatics at Emory University.

ENGAGEMENT WITH INDUSTRY

Researchers in the IDH community are building an expansive network of partners to meet the demand for novel computational solutions that can create competitive advantage in industry. Partners benefit from a diverse and experienced number of Georgia Tech research teams working with IDH that leverage their combined talents to implement state-of-the-art Big Data and HPC solutions. The following portfolio includes a sampling of researchers' collaborations with leading federal agencies and multiple corporations.

Federal Agencies and National Labs

Defense Advanced Research Projects Agency
Department of the Interior
Lawrence Livermore National Laboratory
National Institutes of Health
National Science Foundation
Oak Ridge National Laboratory
Pacific Northwest National Laboratory
Sandia National Laboratories

Industry and Research Foundations

Microsoft
intel
ExxonMobil
Google
NVIDIA
LOGICBLOX®
IBM
BSF United States – Israel Binational Science Foundation
Y!
Reservoir Labs
CRAY THE SUPERCOMPUTER COMPANY

Georgia Tech Spinoff Makes Data Mining Scalable for Business

A Georgia Tech Research Team developing the first open-source, scalable and comprehensive machine learning library, called MLPACK, made a beta release available last year. Created through the FASTLab, directed by **Alex Gray**, Associate Professor in Computational Science and Engineering, the library is currently used by researchers and scientists around the globe in areas such as computer vision, astronomy, computational biology, and computer architecture.

SkyTree, Inc., co-founded by Gray, has now developed a commercialized version of the open-source software. The Skytree Server is the first general purpose machine learning and advanced analytics system, designed to accurately process massive datasets at high speeds. Skytree is designed to connect quickly and easily with existing business IT infrastructure and can be configured to accept data streams from multiple sources and compute near-instantaneous results on each one.

National Leadership in Technology Fields

The National Modeling and Simulation Coalition held an inaugural congress in February 2012 to establish a national agenda for maintaining the growth of modeling and simulation technology and its incorporation into the nation's economy, welfare, and security. **Richard Fujimoto,** Georgia Tech's Director of IDH, is a member of the interim Board of Directors and interim chair of the Education and Professional Development Standing Committee. Fujimoto chaired the inaugural meeting of the committee, which will focus on the entire education pipeline.

The Graph 500 list debuted in 2010 to establish a set of large-scale benchmarks for supercomputing applications and address the need for better hardware architectures for these applications. It ranks supercomputers based on performance of data-intensive applications in various domain and technical areas. **David Bader**, Georgia Tech's Executive Director of HPC and a co-founder of Graph 500, leads the 50-member steering committee of HPC experts in academia, industry and national laboratories that establishes benchmarks for the biannual list.

PARTNERSHIPS

IDH's strategic networks are built on foundational relationships among research experts within the regional innovation ecosystem. Three growing enterprises are highlighted here.

GTRI

As a partner in the Georgia Tech Research Institute's Big Data, Analytics and High Performance Computing (HPC) Strategic Initiative, IDH is working with GTRI to build a core group of researchers to help define and advance the institute's Big Data initiatives. The initial group will use its knowledge base to define what Big Data problems look like, work with applied research projects to identify needed solutions and help grow the campus group into a leading network to solve problems in the space.

Emory Biomedical Collaboration

IDH has partnered with Emory University to create a set of dynamic teams to investigate important medical breakthroughs using state-of-the-art computational tools. The current research projects are an initial collaborative effort to create a broad Biomedical Partnership between Georgia Tech and Emory University in the data and high performance computing fields. The synergy created through these collaborations are shaping the research pathways necessary to advance biomedicine.

Materials Science

A developing partnership between IDH and Materials Science researchers is allowing technical experts to help materials specialists gain access to data and create tools that will be part of a hub for open-source innovation. The envisioned Big Data hub for the Materials research community is a model that, once completed, IDH will actively seek to replicate in order to leverage experts to innovate using computational techniques.

EVENTS AND TRAINING

Research-focused events and training through IDH are a major resource for creating the Big Data culture at Georgia Tech. They enable knowledge transfer to targeted domain areas and create impact by helping researchers identify real-world problems and providing solutions with current methodologies and tools.

IDH Big Data Research Meeting

The Big Data Research Planning Meeting hosted by IDH assembled researchers to discuss initiatives to carry Georgia Tech forward in the burgeoning field. Researchers presented ideas related to the federal "Big Data Research and Development Initiative" and Georgia Tech's positioning to compete for the awards and extending leadership in the area.

IDH Town Halls

In expanding Big Data and HPC initiatives, IDH has held open forums to highlight its role and the resources that are being created to support research and activities in Big Data. The Town Halls also serve to build the networks and collaborations for technical and domain experts.

Workshop on Internet Topology and Economics

The Algorithms and Randomness Center workshop, with sponsorship from IDH, examined Autonomous Systems—thousands of interconnected small networks—that shape the Internet. A diverse group of practitioners examined the networks' economic impact and their future roles.

Workshop on Materials Informatics

Organized by IDH, this workshop drew researchers from across campus in the emerging materials informatics area. It brought together a diverse group of researchers with common interests at the intersection of Big Data, materials science and engineering.

Computational Science and Engineering Seminars

A weekly series of seminars featuring experts in high-performance computing, modeling and simulation, numerical computing, bioinformatics, computational biology, large-scale data analysis and visualization, among other areas. All seminars may be viewed at **smartech.gatech.edu**.

CRUISE Program

The Computing Research Undergraduate Intern Summer Experience is a 10-week program offered to encourage students to consider doctoral studies in Computational Science and Engineering (CSE). The program hosts interns who work with faculty in on-going research projects in CSE.

Virtual Summer School

Georgia Tech was a host site for the course in "Proven Algorithmic Techniques for Many-Core Processors," which was streamed to 11 sites with lectures being delivered through the University of Illinois. The IDH-sponsored event discussed algorithm design choices and impact on HPC.

GPU Workshops

IDH hosted several one-day workshops on Graphics Processing Unit programming, which provided insight into GPU hardware configurations, example codes and references and gave attendees basic knowledge to start using GPUs.

Research Community Takes Leading Position at Annual Supercomputing Conference

Georgia Tech leverages its strengths in high performance computing and Big Data at the annual Supercomputing Conference (SC)—the premier international conference for HPC, networking, storage and analysis—through a unified effort between IDH, the Center for HPC and the Office of Information Technology. Georgia Tech displays the breadth of its research at an exhibition booth and through communications projects on accepted research in the technical program. Faculty and graduate students lead or take part in multiple activities within the technical program at SC, which includes one of the largest industry tradeshows Georgia Tech attends. An average of 40 faculty, students and staff attend the event, with IDH funding and co-managing the exhibition display, organized by **David Bader**, Executive Director for HPC. Georgia Tech's booth provides space for meetings for research groups, potential collaborators and sponsors. It includes a continuous presentation on four monitors of more than 60 current research initiatives in which faculty members are involved and a live feed to the Georgia Aquarium to demonstrate network capabilities through the central campus. Research highlights the comprehensive and interdisciplinary approach at Georgia Tech in tackling a broad range of application areas in massive data and high performance computing.

News and Marketing Development

IDH has created new support systems to better develop research communications that show the context and significance of Big Data- and HPC-driven projects at Georgia Tech. News and marketing platforms have grown in order to communicate results of research, sponsored events, and the larger Georgia Tech growth and activity in these computing fields. IDH develops regular and original news on Big Data research that is targeted to numerous stakeholders. The communications program oversees a web portal for news and events, publications development, tradeshow planning and content creation, among other activities. More details are available at **idh.gatech.edu**.

IDH is exploring the development of a self-perpetuating ecosystem of software tools that benefit research in targeted application areas and create necessary growth for computational solutions. This specialized technical role in Big Data and HPC will better position the Georgia Tech research community in scaling and optimizing computing-related work. IDH also supports infrastructure needs for several research initiatives, notably Keeneland, the leading GPU-based supercomputing project for scientific research through the National Science Foundation.

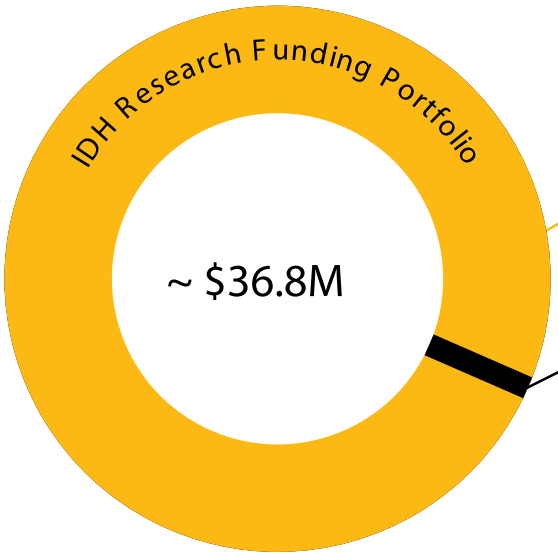
Keeneland Project Deploys Full-Scale System for NSF

In November 2012, Georgia Tech and its partners announced the completed installation and acceptance of the Keeneland Full Scale System (KFS), a supercomputing system available to the National Science Foundation (NSF) scientific community and designed to meet the compute-intensive needs of a wide range of applications through the use of NVIDIA GPU technology. KFS is the most powerful GPU supercomputer available for research through NSF's Extreme Science and Engineering Discovery Environment (XSEDE) program and has delivered sustained performance of over a quarter of a PetaFLOP (one quadrillion calculations per second) in initial testing.

"Many users are running production science applications on GPUs with performance that would not be possible on other systems," says **Jeffrey Vetter**, Principal Investigator and Project Director, with a joint appointment to Georgia Tech's College of Computing and Oak Ridge National Laboratory.

Significant demand for the research system exists, with time-allocation requests already outstripping the total available time for the lifecycle of Keeneland. The Keeneland Initial Delivery system hosted more than 130 projects and 200 users over the past two years.

Georgia Tech's partners on Keeneland include the University of Tennessee-Knoxville, the National Institute for Computational Sciences and Oak Ridge National Laboratory, where the system is housed. More details may be found at **keeneland.gatech.edu**.



Institute for Data and High Performance Computing
Total Funding: $36,862,496

Federal Funding
95%.....$34,867,405

- NSF (National Science Foundation)
  57.5%.....$20,059,176
- DARPA (Defense Advanced Research Projects Agency)
  37.6%.....$13,112,478
- National Institutes of Health, Department of Defense, other agencies
  4.9%.....$1,695,751

Industrial Grants and Contracts
5%.....$1,995,091

Georgia Tech Institute for Data & High Performance Computing

Cover Photo: Argonne National Laboratory

Georgia Tech

Institute for Data & High Performance Computing

Institute for Data and High Performance Computing
Georgia Institute of Technology
Klaus Advanced Computing Building
266 Ferst Drive NW
Atlanta, GA 30332-0280
404.385.4785 • Fax: 404.385.7337
idh.gatech.edu