Cyberinfrastructure at the National Nanotechnology Coordinated Infrastructure (NNCI) Site at Georgia Institute of Technology

The National Nanotechnology Coordinated Infrastructure (NNCI) is an NSF-funded program comprised of 16 sites, located in 17 states and involving 29 universities and other partners. This national network provides researchers from academia, government, and industry with access to university user facilities with leading edge fabrication and characterization tools, instrumentation and expertise within all disciplines of nanoscale science, engineering, and technology. Research undertaken within NNCI facilities is incredibly broad, with applications in electronics, materials, biomedicine, energy, geosciences, environmental sciences, consumer products, and many more. The toolsets of sites are designed to accommodate explorations that span the continuum from materials and processes through devices and systems. There are micro/nanofabrication tools, used in cleanroom environments, as well as extensive characterization capabilities to provide resources for both top-down and bottom-up approaches to nanoscale science and engineering. For more information about NNCI, please visit [www.nnci.net](http://www.nnci.net). Georgia Tech serves as the coordinating office for the NNCI.

Modeling and simulation play a key role in enhancing nanoscale fabrication and characterization as they guide experimental research, reduce the required number of trial and error iterations, and enable more in-depth interpretations of the characterization results. Various NNCI sites provide a diverse set of software and hardware resources and capabilities. Some of these resources are only available to internal users and some to academic users and some to all interested parties. The rest of this white paper describes the rational behind a major cyberinfrastructure at Georgia Tech and its features and capabilities. This computing resource currently serves only students and faculty at Georgia Tech and is not available for external users.

Science and engineering research is the key to understanding everything in our universe and the best way we can improve the human condition. We are on the cusp of answering fundamental questions in the physical sciences, life sciences, social sciences, and mathematical and computational sciences. As our understanding deepens, we can leverage our basic fundamental knowledge to develop innovative and creative technologies that help drive solutions to the most pressing global problems all enabled by advances in cyberinfrastructure.

Investment in heterogeneous, sustainable, scalable, secure, and compliant cyberinfrastructure is critical to enable future discoveries. Significant resources are needed to address the storage, network bandwidth, and massive computational power required for simulation and modeling across multiple scales. Data-centric computing is also vital, necessitating high-throughput analysis and mining of massive datasets, as well as the ongoing demand for low cost, long-term, reliable storage. Sustained investment in cybersecurity will support sharing of datasets along with greater multi-institution and multi-disciplinary research collaboration. A significant investment in software engineering will enable researchers to leverage the promise offered by public-private, multi-cloud based cyberinfrastructure and emerging new architectures. Some of the greatest risks are an inability to meet workforce demand and the lack of a sustainable funding model. Addressing these issues includes maximizing the steady pipeline of students entering science and engineering careers; creating professional retooling programs; building specialized local and regional teams; and leveraging a range of investment sources including federal, state, municipal and local entities, as well as public-private partnerships (e.g. academic and industry, government and corporate).

Future breakthroughs are reliant on continued investment of national level resources in the path to exascale systems. That said, there are real limitations in an approach that primarily relies on "big iron" systems. More broadly, the perception is a general lack of resources to accommodate large simulations due to smaller jobs that require high-throughput computing. This problem is not likely to be addressed by reaching exascale capacity as there is essentially unbounded demand yet natural boundaries to scalability at many levels. Few researchers have access to funding to port code to new architecture introduced by these “big iron” systems. The national scale resources are also not well suited for small to medium-sized jobs and local institutional support is uneven and inconsistent.

Our existing cyberinfrastructure is also limiting for researchers who need more data-centric systems. Many modern computational tasks are "embarrassingly parallel" and have strong scalability, but available computer clusters and HPC systems are not designed or optimized for such HTC workloads. Examples include data analytics and deep learning workloads. We must develop new systems that can more efficiently support data intensive applications. There are promising technologies for this including modern memory hierarchies, GPUs, and other heterogeneous environments.

In 2009, Georgia Tech created a technology model for central hosting of computing resources that would be capable of supporting multiple science disciplines with shared resources, private resources, and a group of expert support personnel, in support of campus research community. This project is called “Partnership for an Advanced Computing Environment (PACE).” Since its inception, PACE has acquired more than 50,000 cores of high performance computing capability and more than 8 Petabytes of total storage used by approximately 3000 (1500 active) faculty and graduate students. This project provides power, cooling, and high-density racks, as well as a three tiered storage system including home directory, project space, and high transfer rate scratch space across the whole system. On top of storage, compute capabilities are provided both as private resources for a researcher or research group, or as a public resource with access open to researchers on campus through a proposal process for requesting compute cycles. PACE is funded through a mix of central and faculty funding that has proven sustainable is expected to continue with increased growth into the future (Figure 1). Due to this rapid growth, more hosting capability is being planned.



Figure 1. Growth at PACE, represented in cumulative terms of CPU Cores in blue, users in pink, faculty in orange, and PACE full-time employees (FTEs) in green.

A significant investment in software engineering will enable researchers to leverage the promise offered by public-private, multi-cloud based cyberinfrastructure and emerging new architectures. Some of the greatest risks are an inability to meet workforce demand and the lack of a sustainable funding model. Addressing these issues includes maximizing the steady pipeline of students entering science and engineering careers; creating professional retooling programs; building specialized local and regional teams; and leveraging a range of investment sources including federal, state, municipal and local entities, as well as public-private partnerships (e.g. academic and industry, government and corporate).