Accelerating Innovation Through Advanced Cyberinfrastructure: A Strategic Vision for Research Cyberinfrastructure at Rutgers

Report edited by:

Helen Berman (SAS)
Manish Parashar (RDI^2/SOE)
Margaret Brennan (ORED)
Nabil Adam (Newark)
Debashish Battacharya (SEBS)
David Foran (RWJMS and CINJ)
Deborah Lazzarino (NJMS)
Michael Lesk (SCI)
Joseph Martin (Camden)
Dimitri Metaxes (SAS)
Les Michaelson (OIT)
Fred Roberts (DIMACS)
Jay Tischfield (RUCDR)

Draft
November, 2013

Accelerating Innovation Through Advanced Cyberinfrastructure: A Strategic Vision for Research Cyberinfrastructure at Rutgers

Executive Summary: Advanced computing and data cyberinfrastructure (ACI) are playing increasingly important roles in all areas of computational and data-enabled science, medicine, engineering, and business. To be internationally competitive, it is critical that Rutgers develop and implement a bold strategic vision for an ACI ecosystem that will provide researchers with essential computing and data handling capabilities, and students with necessary exposure and training. This document calls for strategic investment, comparable to those being made by peer institutions, to drive innovation, improve research capabilities and productivity, and enhance faculty competitiveness. The anticipated benefits of implementing the recommendations outlined in detail below are substantial increases in the scale and impact of science and engineering research at Rutgers.

These recommendations are as follows:

- **Deploy a Balanced Nationally Competitive Advanced Cyberinfrastructure at Rutgers:** Make balanced infrastructure investments in computing, mass storage, and high speed/bandwidth digital communication to provide state-of-the-art capacities and capabilities that can give Rutgers researchers the competitive advantage among Big Ten peer institutions and beyond. Making concurrent complementary investments in more experimental aspects of ACI are required to enable faculty leadership in computing and sustain global competitiveness.
- **Recruit Computational and Data Competencies:** Recruit necessary expertise (both systems and computational) and associated support structures, including cross-disciplinary leadership, faculty lines in computational and computing sciences and engineering, and full time computational and data researchers, educators, and programmers.
- **Establish Multidisciplinary Research and Educational Structures:** Establish multidisciplinary computational and data research structures and effective leadership that will enable integration of research, education and infrastructure; encourage synergistic cross-disciplinary collaborations; ensure curriculum development and provision of learning opportunities; and foster centers of excellence in computational and data-enabled science, medicine, engineering, and business.
- **Establish an Office for Research Cyberinfrastructure at Rutgers:** Rutgers must establish an Office of Research Cyberinfrastructure, lead by a nationally recognized leader in computation and data, that can provide strategic leadership in developing the required ACI at the university, and can coordinate the investments in advanced research infrastructure and expertise necessary to empower research, learning, societal engagement, and competitive advantage to the Rutgers community.

I. Overview

Advanced computing and data research cyberinfrastructure (ACI) (including compute, storage, communication, and expertise) are playing increasingly central roles in all areas of computational and data-enabled research, and are driving innovation and insights in areas such as health and bio-informatics, personalized or precision medicine and drug discovery, understanding social networks and human behavior, advanced manufacturing, transportation, energy, materials, etc. Significant investment in ACI at other universities and research centers within the US and

globally is already underway. To be competitive, it is critical that Rutgers develop a bold strategic plan for establishing a research cyberinfrastructure ecosystem that can provide researchers with essential computing and data capabilities, and students with the necessary exposure and training. Moreover, this plan must be aligned with national research and educational priorities, and augment support for high-quality computational enabled research efforts already in place at Rutgers.

This document outlines such a strategic plan, and is based on insights obtained from multiple systematic surveys[1]: (1) an inventory of the ACI capabilities of key computational researchers; (2) a university wide survey of computational enabled research and associated ACI requirements; and (3) two surveys of faculty concerning specific research opportunities and the deleterious impact of inadequate ACI capabilities and expertise.

Herein, we also provide discussion and examples of computational and data-enabled research programs that are dependent on advanced ACI; missed opportunities and lost funding due to the lack thereof at Rutgers; and productivity losses due to the lack of support and/or expertise where the researchers have been forced to build/maintain ACI to support their research.

This report proposes the establishment of an integrated and robust ACI for research. Such an ACI would interoperate with the common aspects of the administrative, medical and instructional computing infrastructure, and would focus on establishing advanced systems for research computing, communication, data storage, analytics and visualization. It would also foster a community of experts in systems, advanced software applications and computational methods and practices. The impact of establishing such an ACI at Rutgers will include:

- Enabling the development and growth of nationally and internationally competitive research programs.
- Dramatically increasing the number of external funding opportunities for which our faculty can qualify and successful compete.
- Reducing outsourcing of ACI-enabled activities
- Facilitating cutting-edge cross-disciplinary and industry-academic research.
- Fostering a unique educational and training environment that will produce a workforce with advanced computational and data expertise.

II. Current State of ACI at Rutgers

A. Computing Infrastructure Landscape at Rutgers

The current Rutgers computing infrastructure landscape can be broadly classified as Administrative, Instructional, Medical and Research cyberinfrastructure:

- Administrative – computing infrastructure supporting accounting and payroll, human resources, admissions, student records, pre award and post award grants management, university websites, and email, etc.
- Instructional – computing infrastructure supporting courses and instructional activities, ranging from computer laboratories to course management, and e-learning, etc.

[1] Summaries of the inventory and surveys are included as appendices.

- Medical – computing infrastructure supporting patient management, billing and electronic medical records, etc.
- Research – cyberinfrastructure supporting advanced computational and data-enabled science, medicine, engineering, and business research.

These four classes of computing infrastructures are very different in their requirements, the nature of their component systems, and their usage modes. For example, the administrative computing infrastructure is typically composed of enterprise-style data centers composed of conventional commodity components and uptimes, reliability and security are the primary concerns. In the case of medical computing infrastructure, compliance and certification are most critical. In contrast, advanced research infrastructure typically incorporates specialized computing, communication, storage and software components, often with experimental components and at very large scales, and run custom software and applications. What are the most critical concerns for this?

B. ACI-enabled Research at Rutgers

There exist several nationally and internationally recognized computational and data-enabled research programs across campuses, disciplines, academic units and center at Rutgers, which critically depend on the availability of ACI including computing, storage, and communication capabilities, and associated expertise. A sampling of related research areas along disciplinary lines is listed below:

- **Medical and Life Sciences:** Genomic studies of Tourette Disorder; microbiology and infectious diseases studies; community databases of 3D protein structures that includes the curation and dissemination of macromolecular structure data; Connexin channel studies using molecular dynamics simulations; development of novel polymer therapeutics for the management of atherosclerosis; studies of Tyro-3, Axl and Mer receptors using docking studies; simulations of biocatalysis, including a novel linear-scaling fully quantum mechanical (QM) force field; exploratory analysis on large longitudinal medical datasets such as the Medicaid Analytics eXtract (MAX) database; high throughput genomics and phylogenomics to discover the flow of genetic information across the tree of life; design of new biological materials with novel properties; macromolecular crowding as it affects protein behavior and drug delivery; mining Electronic Medical Records to determine and compare patient characteristics with outcomes of cancer care off and on protocol; interrogation and complex queries and stratification of correlated clinical, genomic and image-based information.

- **Physical Sciences and Engineering:** Computational gas dynamics – realistic gas simulations in hypersonic flows, Large Eddy Simulations of shock wave-boundary layer interactions; computational materials science – designing materials with novel properties, long-term stability of alpha plutonium, and studies of hemoglobin, an Fe metallo protein central to human respiration that bind various diatomic ligands (i.e., oxygen and carbon monoxide); environmental networks – environmental observation/modeling/management systems providing streaming 4-D data enabling advanced environmental projections which coupled with human infrastructure models to facilitate the analysis of impacts associated with current and future environmental conditions; earth system modeling – analysis coral larval dispersal in the Indonesian throughflow, multi-scale climate simulations in the California current, fish and fishing fleets in the north Pacific and downscaled climate simulations in the northwest Atlantic; probing the nature of dark energy – using an array of integral field spectrographs to discover over one million new galaxies, and analysis of the three-dimensional spatial clustering of these galaxies to determine the nature of the mysterious dark energy that is

causing the expansion of the universe to accelerate; understanding natural disasters – computational disaster management using large-scale pre- and post- 3D disaster information to create an immersive environment for visualizing damage data in very high resolution to determine damage mechanism and develop remote damage assessment and debris quantification approaches.

- **Social and Library Sciences and Humanities:** Processing of massive text collections to extract meaning, and levels of confidence; social media data mining, information retrieval evaluations; visual exploration of human objects and concepts, exploring the inter-relationships of humanities objects and concepts.

C. Impacts of Inadequate ACI Capabilities

The lack of adequate ACI capacity, capabilities, and expertise is evident, both at the level of individual researchers and across the university as a whole. A statistically-significant association between ACI and research productivity (in terms of publication numbers, impact factors, and funding) has been well documented[2]. Newer empirical research demonstrates that university investment in ACI systems increases both the number of papers published there and institutional NSF funding.[3] Specifically, Apon et al. found that universities that increased their Top 500 rankings (www.top500.org) by one point saw an average increase of 60 research publications in the same year.[4] The regression analysis also revealed that a one point increase in a university's cumulative Top500 ranking score is accompanied by an average concurrent NSF funding increase of $2.4M.[5]

As a specific example, the Pervasive Technology Institute (PTI) at Indiana University, a member of the Big Ten and Committee on Institutional Cooperation (CIC), was a result of a $30 Million investment in 1999, and has since resulted in about $67M in federal grant funding, of which about $24 million was F&A. PTI employs about 120 CI personnel, 54 of whom are grant funded, provides compute and data resources, and houses four research centers. In Spring 2013, it installed a petaflop (10^{15} floating point operations per second) supercomputer, which then was the largest such system that is owned by and operated solely for the benefit of a single university.

Below we provide two specific examples of the research and financial impact of ACI limitations on leading research projects at Rutgers:

Protein Data Bank (PDB): The Protein Data Bank was established in 1971 as the single archive for information about biological macromolecules. In 1998 Rutgers University was funded by the National Science Foundation, the National Institutes of Health and the Department of Energy to lead and manage that resource. The resource is used by academic researchers and educators, as

[2] Amy Apon et al., "Computation and data is playing an increasingly important role in all areas of science and engineering," Journal of Information Technology Impact, Vol. 10, No. 2, pp. 87-98, 2010. (http://www.jiti.net/v10/jiti.v10n2.087-098.pdf).
[3] Amy Apon et al., "High Performance Computing Instrumentation and Research productivity in U.S. Universities," *Journal of Information Technology Impact,* Vol. 10, No.2, 87-98.
[4] Amy Apon et al., "High Performance Computing Instrumentation and Research productivity in U.S. Universities," *Journal of Information Technology Impact,* Vol. 10, No.2, p. 95.
[5] Amy Apon et al., "High Performance Computing Instrumentation and Research productivity in U.S. Universities," *Journal of Information Technology Impact,* Vol. 10, No.2, p. 94.

well as by pharmaceutical scientists involved in drug discovery. It requires 24/7 operations as well as expert data management, storage and fast networking. The PDB has a very large international user community. For example, in 2012, the site had more than 250,000 unique visitors per month, and more than 350,000,000 downloads of data. Although management of the project and most aspects of the work are done at Rutgers, the current Rutgers networking would not allow us to have a robust data distribution system. A strategic partnership was formed with the University of California San Diego; UCSD has two supercomputer centers and very strong network presence that would allow us to accommodate the very large global community of users. The subcontract to UCSD is for about $2,000,000 per year of which $600,000 is for indirect costs. *Over a ten-year period UCSD administration has received more than $6,000,000 for providing this service, over which, we have had no control.* Clearly it would have been to our advantage from both from the public relations perspective, as well as financially, to be able to run the entire operation at Rutgers.

Rutgers University Cell and DNA Repository (RUCDR): For the past six years, Rutgers University Cell and DNA Repository (RUCDR) has outsourced its advanced computational needs to the Information Sciences Institute (ISI) at the University of Southern California. As part of the most recent five-year, $45 M award for the NIMH Center for Collaborative Genomics Research on Mental Disorders (J. Tischfield, PI), RU subcontracts $850,000 per year ($4.3M over the next five years) to ISI for the development of computational tools accessible through the web. Efforts were made to find a partner at RU but no group was willing or able to provide this service. The main issue is that there are no groups at RU that provide service in response to the relatively circumscribed computational research demands of faculty in diverse fields. To some extent, RUCDR has built limited computational resources using its own IT staff and consultants. However, it won't be possible for RUCDR to compete with large computational groups at institutions such as Harvard, Johns Hopkins and UCSD until it can access advanced computational resources as needed. For the past 15 years RUCDR has been outsourcing its medical informatics needs to Washington University School of Medicine (WU), mainly for its NIMH, NIDA and NIAAA grants and contracts. This subcontract to WU has totaled nearly $10M over this period. Even after integration of the medical schools, RU has no presence in medical informatics as it relates to computational genomics. In particular, neither RWJMS nor NJMS have departments of human genetics or centers for computational services.

Key dimensions of the impacts of current ACI limitations include:

- **Research and Education:** The lack of adequate computational, data, and communications capabilities and capacity is significantly impacting computational and data-enabled scientific, medical, engineering and business research at Rutgers. Impacts include limiting the type and scale of research, inability to compete with external peers, and inability to respond competitively to many funding solicitations. The lack of locally available ACI has also limited the level and the type of exposure to computation and data with associated concepts and technologies available to students. This reality is of serious concern, as computation and data are important in all areas of science and engineering, and should be an integral part of curriculum and training.
- **Faculty Productivity:** The lack of readily available expertise and support structures has impacted faculty research productivity. All too often, faculty or students are reduced to managing their own ACI, which is neither appropriate nor efficient, particularly in the face of growing scale and complexity. This deficiency also limits the type and scale of ACI that can be installed and managed.

- **Faculty/Student Recruiting:** Limited ACI and lack of critical mass in computational/data-related research areas is negatively impacting our ability to attract (and retain) computational faculty, researchers, and students across Rutgers. Isolated pockets of excellence that have independently invested in ACI remain the exception. Given the increasing importance of computation and data to all aspects of science, engineering, and society, this trend has ominous implications for the quality of research at Rutgers.

III. Findings and Recommendations

Computational and data enabled research is a central theme cutting across the national research and education agenda, and it is critically important that Rutgers University build core competency in this area. The importance and immediacy of having adequate cyberinfrastructure at Rutgers is further accentuated by the growing role of computation and data in all areas of science, medicine, engineering, and business, and the current and future trends in cyberinfrastructure, such as disruptive hardware innovations, ever-increasing data volumes, complex application structures and behaviors, and new first-order concerns, such as energy efficiency. These trends reflect the continued quest towards extreme scales in computing and data, necessary for innovation in science, medicine, engineering, and business.

Investments in Advanced Research Cyberinfrastructure (ACI): ACI needs permeate all aspects of competitive computational and data enabled research activities, including facilities for high performance computing and communications, data management, advanced visualization, etc. In addition to investing to realize a balanced ACI (compute, storage and communication) comparable to, but ideally better than, peer institutions, it is equally important to invest in the necessary expertise (both systems and computational) and support structures. Our goal should be to support scientific, medical, engineering and business research enabled by ACI, and the science and engineering underpinning new advances in ACI. Such an ACI should also provide researchers with global linkages to national and international cyberinfrastructure resources that will connect Rutgers with observational instruments, data streams, experimental tools, simulation systems, and individuals distributed across the globe.

An equally important area for investment is *experimental cyberinfrastructure*. We need to provide researchers access to leading-edge technologies and/or unconventional design points, addressing emerging concerns such as power-efficiency and resilience, and investigating application-specific system configurations. Examples include hybrid many-core and accelerator based systems, low-power systems, deep memory based systems, and new network technologies and architectures. These systems will enable faculty to keep pace with innovations in computing and computation, increase competitiveness, and provide students with exposure to emerging technologies.

- **Recommendation: Deploy a Balanced Nationally Competitive Advanced Cyberinfrastructure at Rutgers.** Make balanced infrastructure investments in computing, mass storage, and high speed/bandwidth digital communication to provide state-of-the-art capacities and capabilities that can give Rutgers researchers the competitive advantage among Big Ten peer institutions and beyond. Making concurrent complementary investments in more experimental aspects of ACI are required to enable faculty leadership in computing and sustain global competitiveness.

Investments in ACI competencies *will be critical at all levels*. These include faculty in computational and computing sciences and engineering, researchers doing computational enabled research, and support personnel with systems and programming expertise.

- **Recommendation: Recruit Computational and Data Competencies.** Recruit the necessary expertise (both systems and computational) and associated support structures, including cross-disciplinary leadership, faculty lines in computational and computing sciences and engineering, and full time computational and data researchers, educators, and programmers.

Research and Educational Structures: Scientific, medical, engineering and business research is becoming increasingly multidisciplinary. Consequently, it is imperative that Rutgers create structures to support and nurture the development of required collaborations and synergies with the goal of catalyzing the necessary *socio-technical* changes in research across all of science and engineering. Computational and data enabled research also requires educational practices to move beyond traditional university curriculum, experiences, and learning opportunities. Appropriate mechanisms must be created to educate and nurture the next generation of computational scientists and engineers by providing them with the necessary foundation/experiences to address grand challenges of science, medicine, engineering, and business using ACI. This objective requires a concerted assembly of necessary expertise in, for example, computational models, algorithms, HPC, data and visualization technologies, software, and multidisciplinary collaborations.

- **Recommendation: Establish Multidisciplinary Research and Educational Structures.** Establish multidisciplinary computational and data research structures and effective leadership that will enable integration of research, education and infrastructure; encourage synergistic cross-disciplinary collaborations; ensure curriculum development and provision of learning opportunities; and foster centers of excellence in computational and data-enabled science, medicine and engineering.

Community of Excellence: Building on adequate CI, Rutgers must create comprehensive and internationally competitive multidisciplinary CDS&E structures that can provide the leadership required to catalyze and nurture the integration of research, education, and infrastructure, and to foster a community of excellence in computational and data enabled research. Such a hub (or hub of hubs) should provide the requisite outreach and linkage to computation-oriented science and engineering units and research centers at Rutgers. Specifically, Rutgers must establish an ***Office of Research Cyberinfrastructure*** that can provide strategic leaderships, and can coordinate investments in advanced infrastructure and expertise aimed at empowering research, learning, and societal engagement and providing a competitive advantage to the Rutgers community. The anticipated impact is a revolutionary wholesale advance in the scale and impact of science and engineering research at Rutgers.

- **Recommendation: Establish an Office for Research Cyberinfrastructure at Rutgers.** Rutgers must establish an Office of Research Cyberinfrastructure headed by a nationally recognized leader in computation and data that can provide strategic leadership, coordinate investments in advanced infrastructure and expertise aimed at empowering research, learning, and societal engagement, and providing competitive advantage to the Rutgers community[6].

[6] An example is the Office of Research Cyber Infrastructure at University of Michigan (http://orci.research.umich.edu), a member of the Big Ten and CIC.