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

Draft Report
June 24, 2014

ACI Strategic Planning Committee:

Overall Leads:

- Helen Berman (SAS), Manish Parashar (RDI^2), Don Smith (OIT), Margaret Brennan-Tonetta (ORED)

Working Group Leads:

- **Medical**: David Foran (RWJMS and CINJ), Deborah Lazzarino, (NJMS)
- **Biological (Life) and Environmental Sciences**: Debashish Bhattacharya (SEBS), Joseph Martin (Camden)
- **Physical and Earth Sciences**: Oscar Schofield (SEBS), Kristjan Haule (SAS)
- **Computing and Information Sciences**: Nabil Adam (Newark), Michael Lesk (SCI)

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 that being made by peer institutions, to drive innovation, improve research capabilities and productivity, and enhance faculty competitiveness. The anticipated benefits of implementing the recommendations detailed below are substantial increases in the scale and impact of science and engineering research at Rutgers, enhanced training opportunities for students, and a significant rise in external funding secured by Rutgers researchers.

The strategic planning process included multiple systematic surveys[1], several focused meetings with faculty, researcher groups, systems staff and administrators, and focused round-tables with the finance and pharmaceutical industry. The surveys included: (1) an inventory of the ACI capabilities of key computational researchers; (2) a university wide survey of computational enabled research and associated ACI requirements; (3) two surveys of faculty concerning specific research opportunities and the deleterious impact of inadequate ACI capabilities and expertise; (4) an overall cyberinfrastructure inventory for administrative, clinical, instructional and research cyberinfrastructure and related services; and (5) a survey aimed at further understanding the ACI needs of the different constituencies. Four working groups,[2] which included almost 100 researchers and systems management staff, were formed to execute this last survey and to ensure that Rutgers constituencies across all disciplines and campuses were represented. Each working group (Medical, Biological and Environmental Sciences, Physical and Earth Sciences, Computing and Information Sciences) produced a report summarizing the capabilities, requirements and recommendations of the different constituencies. These reports can be found in the Appendix.

Recommendations:

- **Establish an Office for Research Cyberinfrastructure at Rutgers:** Rutgers must establish an Office of Research Cyberinfrastructure that is led by a nationally recognized leader in computation and data. This office will provide strategic leadership in developing the required ACI at the university, and will coordinate the investments in advanced research cyberinfrastructure and technical expertise necessary to empower research, learning, and societal engagement. The leadership provided by the Office of Research Cyberinfrastructure, and the computational and human resources it secures, will give Rutgers University to build its competitive advantage.
- **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. Concurrent, complementary investments in more experimental aspects of ACI are also required to enable faculty leadership in computing and sustain global competitiveness.

[1] Summaries of the inventory and surveys are included as appendices and are also available at http://rdi2.rutgers.edu/ruaci/.
[2] Members of each working group are listed in Section VI.

- **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 data-enabled science and engineering, computer science 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 the 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.

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, the understanding of 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[3], several focused meetings with faculty, researcher groups, systems staff and administrators, as well as focused round-tables with the finance and pharmaceutical industry. The surveys included: (1) an inventory of the ACI capabilities of key computational researchers; (2) a university wide survey of computational enabled research and associated ACI requirements; (3) two surveys of faculty concerning specific research opportunities and the deleterious impact of inadequate ACI capabilities and expertise; (4) an overall cyberinfrastructure inventory for administrative, clinical, instructional and research cyberinfrastructure and related services; and (5) a survey aimed at further understanding the ACI needs of the different constituencies. Four working groups were formed to execute this last survey and to ensure that Rutgers constituencies across all disciplines and campuses were represented. Overall, the working groups included almost 100 researchers and system management staff.[4] Each working group (Medical, Biological and Environmental Sciences, Physical and Earth Sciences, Computing and Information Sciences) produced a report summarizing the capabilities, requirements and recommendations of the different constituencies. These reports can be found in the Appendix.

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 the 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. Understanding ACI Capabilities, Opportunities and Requirements at Rutgers: The ACI Strategic Planning Process

[3] Summaries of the inventory and surveys are included as appendices and are available at http://rdi2.rutgers.edu/ruaci/.
[4] Members of each working group are listed in Section VI.

Efforts to create an ACI strategic plan began in Fall 2012. Three systematic faculty and staff surveys were conducted. The first was an inventory of the cyberinfrastructure capabilities of key computational researchers in which the specific resources used by these groups were identified. The second was a university wide survey of computer enabled research and associated cyberinfrastructure requirements. Finally there was a focused survey of key CDS&E faculty concerning specific research opportunities and the deleterious impact of inadequate CI capabilities and expertise.

Focused meetings with faculty, researcher groups, systems staff and administrators, as well as focused round-tables with the finance and pharmaceutical industry followed these surveys. In February 2013, a report entitled "Accelerating Innovation Through Advanced Cyberinfrastructure: The Urgent Need for Strategic Planning for Cyberinfrastructure at Rutgers" was submitted to the Vice President of Research. A presentation to President Barchi was made in June 2013. Strategic planning meetings were held during the summer of 2013 and a strategic vision report was completed in November 2013.

In the next phase of the ACI strategic planning process an overall cyberinfrastructure inventory for administrative, clinical, instructional and research cyberinfrastructure and related services was compiled in order to provide a context for understanding advanced research cyberinfrastructure (ACI). Four working groups, viz., Medical, Biological and Environmental Sciences, Physical and Earth Sciences, Computing and Information Sciences, were formed representing the different areas of research across all Rutgers campuses in New Brunswick, Newark and Camden. A final survey aimed at further understanding the ACI needs of the different constituencies was prepared, executed and analyzed. Each working group produced reports that summarized the capabilities, requirements and recommendations of the different constituencies. A retreat was held on May 1, 2014 at which the four working groups reported on their findings to the Senior Vice President of Research and Economic Development.

III. 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.
- 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 that are comprised of conventional commodity components and uptimes. Managers of these systems are primarily concerned with issues of reliability and security. 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 runs custom software and applications.

The results of our surveys indicate that Rutgers has only modest computing and storage capabilities, largely in the form of clusters owned by individual researchers or research teams. (Full survey results are available in the appendix.) Aggregate capacity at Rutgers is ~26,000 computing cores and ~1,300 Terabytes of storage. This capacity is decidedly inadequate when compared to individual systems at top research universities. The Stampede system at the University of Texas at Austin, for instance, encompasses 102,400 cores and the Big Red II system at Indiana University will alone have 21,824 cores. Examples from CIC/Big 10 institutions include the recently announced DeepThought2 system at the University of Maryland and the Conte Cluster at Purdue University, which contain 209,200 and 77,520 cores, respectively.

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. Some examples include:

The Protein Data Bank (PDB) is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. PDB structures of macromolecules and macromolecular–ligand complexes provide direct experimental insights into protein function. These structures provide essential information for the understanding of biochemical processes and are critical data for structure-based drug design studies. The Research Collaboratory for Structural Bioinformatics (RCSB, www.rcsb.org/pdb) has managed the PDB in the US since 1998 providing delivery sites at both Rutgers and UCSD. In addition to the PDB, the Berman group in the Proteomics Center develops and maintains the specialized structural biology databases: the Protein Structure Initiative Structural Biology Knowledgebase (sbkb.org), the electron microscopy repository EMDatabank.org, and the Nucleic Acid Database (ndbserver.rutgers.edu). The NSF, NIH, and DOE at the level of $9M/year provide funding for these resources.

The High-throughput Genomics Facility that is run within the Department of Genetics and Human Genetics Institute of NJ and RUCDR Infinite Biologics, and is a public-private alliance. The facility houses the NIMH Center for Collaborative Genomics Studies on Mental Disorders, the NIDA Center, two NIAAA project centers, and the provider for a CINJ Clinical Diagnostics Collaboration. It is also involved with many projects at RU (e.g., Autism and Tourette), the Million Veterans Program, and the Bioprocessing Alliance, which includes Glaxo SmithKline and Merck, among others. These projects require either Nextgen DNA sequencing or genomic microarrays which generate hundreds of terabytes of data and many hours of computation on a cluster with ~1000 cores and 7000 GB of RAM. Many institutions have ported these activities to "supercomputers." The work is supported by the NIH, the Simons Foundation, Glaxo SmithKline, and private gifts that amount to ca. $25 million per year. In addition to completing these funded projects, RUCDR Infinite Biologics provides services for many SAS, SEBS and RBHS faculty members.

The NeuroProteomics Core Facility within the Center for Advanced Proteomics Research (CAPR) provides researchers from the National Institute of Neurological Disorders and Stroke (NINDS) with access to advanced proteomics technologies, and serves as a hub for collaborations and new technology developments. New funding is providing NINDS PI's with access to the latest state-of-the- art technologies with upgraded Core equipment. These expanded capabilities include an Applied Biosystems 4800 tandem mass spectrometer and Thermo Orbitrap and Q Exactive LC/MS/MS systems for high throughput protein biomarker discovery and validation. Critical functions that are enabled by these technology upgrades include: 1) more sensitive identification and 2) quantification of protein phosphorylation sites and other post-translational modifications. The systems significantly enhance the productivity of more than eleven neuroscience research projects (including over 30 NINDS investigators). Mass spectrometry data is “big data” that requires advanced computer infrastructure in the form of computational power and storage.

The Genome Cooperative (GC) was established in 2009 and represents a growing group of environmental and life science researchers who use GC resources and expertise to advance genome research at SEBS and more broadly at Rutgers. This initiative is powered by two *Illumina* genome sequencers (GAIIx, MiSeq) and computational facilities associated with the Bhattacharya lab and at other locations on campus. Over 20 research groups at Rutgers, New Brunswick and eight other US and international groups work with the GC to generate and analyze genome and functional genomics data. These projects are funded by the NSF, DOE, the Gordon and Betty Moore Foundation, and other external sources and have been underwritten by the NJAES.

The Mid-Atlantic Regional Association Ocean Observing System, which is comprised of Distributed, integrated observational/modeling networks. Rutgers is a leader in real-time ocean observation networks and ocean prediction. These advances are changing traditional approaches and the resulting changes are demonstrating significant value to society through improved exploration, discovery, and construction of real-time, three-dimensional maps aggregated through wireless networks. This new vision of environmental omnipresence will expand scientific understanding and improve predictive models of the physics, chemistry, biology and economics of coastal systems. The ocean observational group currently maintains a central capability that includes 17 Linux servers (including ocean modeling), 2 satellite ground stations, 2 9-screen visualization walls, 17 remote site systems (not including autonomous robotic systems), and required storage on the order of 70 TB. Continued leadership in this area will require the acquisition of an advanced computing environment that includes 1) extensive parallel computing that allows for numerical simulations of the ocean and atmosphere to be run with sufficient speed to provide operational significance, 2) redundant robust cyber-backbone which allows operational data streams to be maintained even during extreme events such as Nor’Easters and Hurricanes when the data has maximum value, and 3) sufficient bandwidth to support large data sets collected from distributed networks and ensemble numerical simulations.

The Condensed Matter Physics group is developing new tools for ab-initio description of properties of materials using advanced state-of-the-art quantum mechanical methods such as Density Functional Theory and Dynamical Mean Field Theory. With no experimental inputs, these methods can predict existence of new materials and their physical properties, such as the conductivity, ferroelectricity, magnetism, superconductivity. The computer infrastructure consists of a Beowulf cluster with over 3000 cores, a combined computational power of 35 TFlops and combined storage of 150 TB. It also uses national computational resources (NSF-XSEDE). To remain competitive the project needs <u>more than</u> 100 million core-hours/year and at least 4GB/core.

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[5]. Newer empirical research demonstrates that university investment in ACI systems increases both the number of papers published there and institutional NSF funding.[6] 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.[7] 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.[8]

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 founded using 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 to University of Indiana researchers, and houses four research centers. In Spring 2013, it installed a petaflop (10^{15} floating point operations per second) supercomputer which at the time was the largest such system to be owned by and operated solely for the benefit of a single university. Other universities such as University of Maryland, University of Florida, and University of Michigan have had similar initiatives.

Below we provide 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 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 cannot support a robust data distribution system. To address this deficiency, a strategic partnership was formed with the University of California San Diego that has two supercomputer centers and a strong network presence that can accommodate the PDB's 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, from both a public relations and financial perspective, it would have been advantageous for Rutgers' to run the entire operation internally.

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

---

[5] 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).

[6] 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.

[7] 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.

[8] 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.

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.

Rutgers Cancer Institute of New Jersey (RCINJ): RCINJ is undertaking several key translational and clinical research and development projects, which will require a standards-based informatics platform that can support both small and large-scale investigations. A cornerstone for these projects will be the clinical data warehouse, which will integrate data arising from Electronic Medical Records (EMR); Clinical Trial Management Systems (CTMS), Tumor Registries, Biospecimen Repositories, Radiology and Pathology archives and next generation sequencing devices. Given the large and growing volume of data and the number of modalities that are actively gathered and archived as part of these investigations, the primary challenges for these undertakings have now become the scarcity of adequate computational and data analytics resources. These applications offer tremendous potential for collaborative projects and research proposals, which transcend basic and clinical research. In order to continue to propel these projects forward, it is essential that we establish the requisite cyber-infrastructure and associated computer and network resources, which allow reliable processing of the rich informational content extracted from large patient cohorts. The capabilities provided by these resources will make it feasible to conduct high-throughput screening and mining of large data sets, and to generate and test hypotheses. These capabilities will enable the community of investigators to stratify patient populations in multi-dimensional space; perform dynamic modeling of the changes in the molecular signatures and morphology; visualize organs, tissues and microstructures in 3D and determine precise localization of biomarkers within the tumor environment throughout the course of disease progression. Together these advances and new technologies will serve to improve prognostic accuracy and therapy planning for subpopulations of patients who have been afflicted with cancer while facilitating investigative cancer research and discovery.

Impacts of current ACI limitations: The Key Dimensions

- **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. Due to the lack of coordinated and available system support staff, all too often, faculty or students are reduced to managing their own ACI. This 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.
- **Financial:** Over the last 15 years, the Rutgers Protein Data Bank and the Rutgers University Cell and DNA Repository have paid the University of California San Diego and Washington University School of Medicine almost $30 million for access to their advanced cyberinfrastructure capabilities. Of these funds, $6,000,000 went directly to UCSD central administration as F&A. This is one example of forgone F&A that could have been used to enhance Rutgers resources instead of those of another institution.

IV. Findings: ACI Requirements and Gaps

The most recent survey from Spring 2014 (Appendix 5) and the ensuing discussions within the ACI strategic planning committee highlighted significant gaps between current ACI capabilities at Rutgers and the ACI capabilities required by research. These gaps spanned all aspects of ACI, and are detailed below.

Centralized Management and Support for ACI: A university-wide entity that could coordinate the acquisition of and access to ACI resources, such as an Office for Research Cyberinfrastructure, was cited as a crosscutting critical need and gap.

Datacenter: The need for a university-wide reliable and state-of-the-art Tier-2 datacenter for housing ACI was clearly highlighted by the survey. Current university-level hosting capabilities are severely limited in capability and capacity and are outdated technologically. Efforts to address similar deficiencies at peer universities have demonstrated the effectiveness and impact of such an investment.

Networking: The survey indicated the need for increasing the backbone network connectivity across campus by 1 to 2 orders of magnitude to at least 10 Gb/s. Furthermore, the need for addressing the "last-mile" and ensuring access to this bandwidth at the desktop was highlighted. Finally, survey respondents cited the need for connectivity to the Internet2 national (and international) high-speed backbone through a regional point of presence in New Jersey.

Computing: The survey indicated a critical need for state-of-the-art computing capabilities on the order of 100's of Teraflops with adequate memory and equipped with high-speed communication backplanes and accelerators (e.g., GPGPUs). Researchers also expressed a desire for seamless integration with national cyberinfrastructure such as NSF's XSEDE network and the Open Science Grid (OSG).

Storage: The need for centrally managed reliable shared storage in the order of 1-10 PB was noted in the survey as a critical need across the different research areas. The data-management and data preservation requirements instituted by funding agencies further accentuated this need. Some groups also noted the need for archiving and remote backup. In the medical and healthcare domains, HIPPA and FISMA compliance, as well as data security and privacy were also cited as critically important. Compliant, secure and reliable storage provided as a service by peer universities to their researchers (e.g., Indiana University) is a model that should be considered at

Rutgers. The need for reliable, ETL interfaces to allow automated input and organization of data from disparate sources including EMR's, genomic archives, PACS was also noted.

Personnel: The surveys highlighted the critical need for multiple types of ACI expertise at Rutgers, including system administrators, application programmers, technical consultants, and computational scientists. The survey results also highlighted the need for consolidating these resources and developing a model for sharing them across research groups. This is consistent with best practices across peer universities. The need for maintaining a directory of relevant expertise was also noted. Focused "cluster hires" in computational and data related areas across all units at Rutgers was cited as essential to position Rutgers as a leader in this important areas.

Training and Outreach: Survey respondents cited the need for university-wide training and outreach including courses, workshops and technical consulting services. The need for (and potential impact) of a crosscutting multidisciplinary curriculum in computational and data sciences was also noted.

V. 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 establish core competency in this area. Such investments should build on and leverage existing areas of strength at Rutgers and should complement existing capabilities. 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 that is necessary for innovation in science, medicine, engineering, and business.

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 leadership, 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, enhanced training opportunities for students, and a significant rise in external funding secured by Rutgers researchers.

- **Recommendation: Establish an Office for Research Cyberinfrastructure at Rutgers.** Rutgers must establish an Office of Research Cyberinfrastructure that is headed by a nationally recognized leader in computation and data. This office will provide strategic leadership, and will coordinate investments in advanced infrastructure and expertise aimed at empowering research, learning, and societal engagement. Ultimately, the leadership the Office of Cyberinfrastructure provides, and the computational and human resources it secures, will allow Rutgers University to build its competitive advantage[9].

[9] 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.

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 data-enabled science and engineering, computer science 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 and associated support structures, including system administrators, application programmers, technical consultants and computational scientists, and develop a model through which these resources can be shared. Focused "cluster hires" in computational and data related areas across all units at Rutgers should also be explored.

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.

VI. Action Items and Timeline

1. **Short term (12 Months):** In the next 1-3 months, effort should focus on developing business and execution plans for establishing the Office of Research Cyberinfrastructure. In the short term, an overarching Rutgers-wide coordination and management structure in the form of the Office of Research Cyberinfrastructure should be established within the next year. Partnerships should be established between this office and existing areas of strength identified in the working group reports.
2. **Intermediate term (1 – 3 Years):** In the near term, activities should include establishing key structures for research and education that can support multidisciplinary computational and data-enable science as well as deploying the ACI core that can support immediate research/education needs. Mechanism for engaging with stakeholders across Rutgers should be defined.
3. **Longer term (3+ Years):** Over the longer term, mechanisms for ensuring sustained investments in ACI and its seamless integration into all aspects of research and education across Rutgers should be developed. Mechanisms for oversight (internal and external) and adaptation/correction should also be established.

VI. ACI Strategic Planning Committee

Leadership Team

Helen Berman (SAS), Manish Parashar (RDI^2), Don Smith (OIT), Margaret Brennan-Tonetta (ORED)

Working Groups

Medical Working Group: David Foran (RWJMS and CINJ), Deborah Lazzarino, (NJMS), Frank Sonnenberg (RWJMS - Ambulatory Medicine), John Kerrigan (CINJ), Robert Wieder (RBHS), Petros Levounis, (NJMS), Bin Tian (RBHS-Newark, NJMS), Stephen Crystal (Rutgers-NB IHHCPAR), Pamela A. Ohman-Strickland, Ph.D. (RBHS-NB-EOHSI), Panos G. Georgopoulos (RBHS-NB-EOHSI), Bill Halperin (RBHS-NJMS-Newark), Vlad Belyi (RWJMS - Sys Bio), Murray Becker (RWJMS - Radiology), Evan Cadoff (RWJMS - Clinical Pathology), Chang Chan (RWJMS - Sys Bio), Lauri Goodell (RWJMS - Pathology), Susan Goodin (RWJMS - Pharm/Clinical Trials), Bruce Hafty (RWJMS - Rad Onc), Thomas Kearney (RWJMS - Surg Onc), John Nosher (RWJMS - Radiology), Gratian Salaru (RWJMS - Pathology), Karen Shapiro (RWJMS - Clinical Operations), Mark Stein (RWJMS - Med Onc), Deborah Toppmeyer (RWJMS - Med Onc), David Weissmann (RWJMS - Pathology), Ning Yue (RWJMS - Rad Onc), Gokce Toruner (NJMS), Michael Sirken(NJMS), Petros Levounis (NJMS), Masayuki Shibata (SHRP), Sayed Haque (SHRP), Gwen Mahon (SHRP), Dennis Carmody (SN), Sally Hodder (NJMS), Soteropoulos (NJMS), Charles Prestigiacomo (NJMS), Bin Tian (NJMS), Betsy Barnes (NJMS), Marila Gennaro (NJMS), Syed Haque (SHRP), Steve Hanson (Newark)

Biological (Life) and Environmental Sciences: Debashish Bhattacharya (SEBS), Joseph Martin (Camden), Jay Tischfield (RUCDR), David Case (BioMaPS), Tara Matisse (SAS), Yana Bromberg (SEBS), Alexander Schliep (SAS), Gyan Bhanot (SAS), Roman Shirikov (SAS), Bin Tian (NJMS), Andrew Harris (NJMS), Vikas (Vik) Nanda (RWJMS), Benedetto Piccoli (CCIB, Camden), Gaetano Montelione (RWJMS), Bart Krekelberg (Newark), James Tepperto (Newark)

Physical and Earth Sciences: Oscar Schofield (SEBS), Kristjan Haule (SAS), Gabriel Kotliar (SAS - Phy), Anthony Broccoli (SEBS), David Vanderbilt (SAS - Phy), Karin Rabe (SAS - Phy), Enrique Cuchitser (SEBS), Liping Liu (SOE - Math), Steve Garofilini (SOE), Rachel Somerville (SAS), Amith Lath (SAS), Robert Kopp (SAS), Rachel Somerville (SAS Phy), Amitabh Lath (SAS Phy), Tai-Sung Lee (SAS Che), Prosenjit Bagchi (SOE), Michele Pavanello (Newark)

Computing and Information Sciences: Nabil Adam (Newark), Michael Lesk (SCI), Dimitri Metaxas (SAS), Les Michaelson (OIT), Fred Roberts (DIMACS), Rich Martin (SAS), Thu Nguyen (SAS), Tomasz Imielinski (SAS), Ping Li (SAS), Tony Zhang (SAS), Jon L. Oliver (SCI), Jim Boyce (NJMS), Patrick Mattis (SN), Matthew Rosenthal (SHRP), Charles Hedrick (OIT), Faizan Ahmed (ORED), Sean Campbell (Athletics), E. A. Elsayed (SOE), Liviu Iftode (SAS)

VI Appendices

1. Report from the working groups
2. Results of the surveys