TO: Cyberinfrastructure Research Taskforce Committee and the Indiana University community

FROM: Craig A. Stewart, Executive Director, Pervasive Technology Institute; Associate Dean, Research Technologies, Office of the Vice President for Information Technology and CIO

DATE: 25 June 2010

SUBJECT: Progress Report on Implementation of Recommendations from the Indiana University Cyberinfrastructure Research Taskforce

Dear Colleagues,

As we continue implementation of *Empowering People: Indiana University's Strategic Plan for Information Technology,* it is appropriate to look back at our progress in fulfilling recommendations made in the report of the 2005 Cyberinfrastructure Research Taskforce (CRT). This taskforce was commissioned in 2004 by then-Vice President of Research Michael McRobbie, and chaired by Bradley C. Wheeler, now Vice President for Information Technology and CIO. The full report of this committee is available online at https://scholarworks.iu.edu/dspace/handle/2022/469.

The summary that follows conveys progress-to-date on the ten recommendations in the CRT final report. Even as we face unprecedented budgetary challenges, IU aims to maintain "excellent facilities for research and education," as stated by President McRobbie. In this memo I summarize some of the key activities undertaken in response to the 2005 Cyberinfrastructure Research Taskforce report by the Office of the Vice President for Information Technology (OVPIT), University Information Technology Services (UITS), and its partners in developing and providing IU's advanced research cyberinfrastructure, particularly the School of Informatics and Computing (SOIC), Digital Library Program (DLP), and Pervasive Technology Institute (PTI). (Note that PTI is itself a collaborative effort of the School of Informatics and Computing, Maurer School of Law, Office of the Vice President for Information Technology, and University Information Technology Services.)

CRT Recommendation #1: *Indiana University should continue investments in core IT infrastructure that is a foundation for IU's advanced cyberinfrastructure. The university should expand the successful principles of equipment life cycle budgeting in line with the ITSP to all levels (schools, departments, etc.) to ensure the long-term sustainability of the core IT infrastructure required by scholars.*

The expansion of life cycle budgeting approaches for all core infrastructure remains a goal, and an increasingly difficult one in tough budget times. A recent review of the life-cycle funding (LCF) program has also assessed how new models of virtualization can provide more sustainable options for some services that had formerly relied on local equipment refreshes. University Information Technology Services (UITS) continues to manage core infrastructure, such as public computing labs and classrooms, with required LCF for any expansion.

The new IU Bloomington Data Center also represents a critical element of IU's core information technology infrastructure. This hardened facility plays a key role in protecting valuable

institutional data, high performance computational and storage systems for research, administrative environments, and key networking assets for Indiana University. It is vital to maintain, expand and provide sustainable funding for this facility in order to meet IU's growing need for the cyberinfrastructure that is critical in supporting IU's mission.

CRT Recommendation #2: *Indiana University should continue its investment in site licenses for software and datasets as scholarly tools for data analysis and interpretation for use on both personal and university-owned workstations. Whenever possible, such licensing arrangements should encompass the entire university.*

UITS has expanded its site licensing for 28 statistical, mathematical, and geographical information system (GIS) software titles. Two more software packages are managed in collaboration with the IUB Libraries. Licensed copies of most software titles are resold to faculty, staff, and students at highly discounted rates (typically $25 to $200). Of the 30 software titles, 24 are available to departments, faculty, and staff on all campuses. Fifteen are also available to students on all campuses. These software licensing efforts helped IU achieve a cost avoidance of $12,012,508 during FY 08/09.

CRT Recommendation #3: *Indiana University should continue to execute and accelerate an incremental and extensible strategy that enhances its overall storage infrastructure from online storage to long term archival storage. Dependable archival storage must include a commitment to ongoing and periodic data validation and maintenance of software for reading and migrating the data to newer formats.*

Indiana University is making strong investments in storage through ongoing upgrades and storage library replacements to our Scholarly Data Archive (which replicates data in Bloomington and Indianapolis, in order to ensure data are securely stored even in face of a disaster). Introducing Linear Tape-Open (LTO) tape capability, along with our standard TS1130 tape capability, will allow us to avoid relying on a single vendor-implemented tape archival storage strategy. IU's Research File System allows access for desktop environments with many different access protocols, allowing for medium performance online disk storage. This system is tied into the Scholarly Data Archive. IU offers very high performance online storage through the Data Capacitor, which is utilized both on our computational environments and across the Wide Area Network. IU is also involved in the HathiTrust digital repository and in leading a pilot test of a research storage service to be shared by institutions of the CIC (Committee on Institutional Cooperation – the Big 10 plus the University of Chicago).

CRT Recommendation #4: *Indiana University should enhance its networks through optimized engineering or capacity growth to include much faster end-to-end network capabilities from specific points of need (laboratories, offices, classrooms) to IU's central computing facilities and national and international research networks.*

UITS has installed several targeted upgrades of networks to support particular research needs in chemistry, astronomy, medicine, and biology. Network connections to national research networks have also been enhanced and total bandwidth expanded. A complete, ten-year network master plan is now being rolled out on the core campuses. Research buildings and lab facilities with very high capacity needs are at the front of the schedule. All buildings will be upgraded to faster and redundant connections to the campus network backbone.

CRT Recommendation #5: *Indiana University should research, develop, acquire, and implement new capabilities for the collection, annotation, and provenance management of data generated by IU researchers. Development of these capabilities should provide for annotation and management of massive streams of data, facilities for metadata management and reusability (such as XML- and standards-based data annotation), and management of data provenance.*

Systems, tools, and processes for dealing with the "data deluge" remain an ongoing area of area of concern for the federal agencies and the research community generally. Data-centric computing is a complex challenge that Indiana University is approaching through research and development related to specific challenges, and through the development of general tools for data collection, annotation, and provenance. Examples of projects both specific and general include:

- UITS has recently completed the development of procedures and risk analyses in order to make it possible for IU researchers to store and analyze electronic protected health information on IU's supercomputers and massive data storage systems. The ability to analyze medical records without having first to de-identify them is of great value to the IU health research community. As of June 2010, IU is one of only two supercomputer centers in the US to have this capability.
- The innovative IU Data Capacitor, developed by UITS and PTI, has established new records in its ability to support the transfer of data over long distances and to manage the input and output of massive streams of data.
- Dr. Beth Plale, professor in the School of Informatics and Computing and Director of the Data to Insight Center (part of the Pervasive Technology Institute) has pioneered tools for the automated recording and management of metadata and provenance management. These tools have been proved in nationally recognized research applications such as the tornado prediction system LEAD (Linked Environments for Atmospheric Discovery). A particularly important new development is the Karma provenance management toolkit. Collecting and maintaining metadata (data describing a particular dataset) and provenance (a record of the ownership, transport, and manipulation over time of a data set) are essential.
- The IU Digital Library Program, a core partner with PTI's Data to Insight Center, continues to investigate utilizing IU mass storage resources to offer faculty end-to-end data-curation services. Early tests enabled dataset registration within the IU ScholarWorks discovery system and access through the Scholarly Data Archive tape storage resources. Recent advanced projects include work with the CIC Research File System Project (Scale Out File System SoFS and GPFS-WAN) and work within the Data to Insight Center in drafting an IU call for proposals to find real-world test implementations that would investigate data-curation for both datasets and database-related content from within a lab workflow system. Work with the CIC Research File System Project has shown the feasibility of using the Fedora digital preservation system within a managed wide-area file system across multiple nodes at CIC institutions. Further work on Fedora is needed to enable data-curation on both IU resources as well as on cloud-based storage resources through the integration of our services with the DuraCloud Project, which is currently in beta test mode. In addition to using external cloud resources we are currently investigating data replication services between the IU ScholarWorks repository and the Texas Advanced Data Center in Austin.

- IU continues to provide hosting service for Flybase, the definitive resource for fruit fly genomics. Since publication of the CRT report, UITS and PTI have aided research in the Department of Biology at IU Bloomington in operating this important data resource, which tracks data and data provenance for research based on the genetics of the fruit fly.
- IU manages the informatics core for the National Institutes of Health-supported Collaborative Initiative on Fetal Alcohol Spectrum Disorder (CIFASD). This international collaboration is working to develop new methods for diagnosing and treating fetal alcohol spectrum disorders – the range of impairments that are suffered by individuals who were exposed to alcohol *in utero* as a result of their mother's drinking. IU has been awarded the critical role of managing and maintaining all of the data collected and used in this multi-institutional research project. We have developed and published a comprehensive data dictionary for fetal alcohol spectrum disorder research, and created a facility that verifies data and manages provenance as researchers from multiple countries enter data into the only shared, comprehensive database for FASD research.
- IU provides the informatics resources for the National Gene Vector Biorepository, an NIH-funded national center that manages gene therapy agents such as stem cell lines for healthcare research.
- The Institute for Digital Arts and Humanities, a partnership of the Office of the Vice Provost for Research, IU Libraries, Digital Library Program, and OVPIT, has supported the development and implementation of data management tools. Most important among these is the EVIA digital archive system, developed under the direction of IU ethnomusicologist Ruth Stone, for management and annotation of digital video.

CRT Recommendation #6: *Indiana University should provide a service for maintaining and publishing of digital datasets within and beyond the university. This service should enable scholars to securely maintain annotation and provenance through appropriate review mechanisms – analogous to journal and conference publication processes used today in the academic community – and provide for the ongoing re-use of IU's scholarly data.*

IU ScholarWorks is a set of digital storage and access services from the Indiana University Libraries and Indiana University Digital Library Program (which includes OVPIT and UITS as partners). A key goal of the IU ScholarWorks project is to make the work of IU scholars freely available and ensures that these resources are preserved and organized for the future. Since the completion of the CRT report, the IU Libraries and IU Digital Library Program have expanded services so that now many disciplines store research papers and reports. Most recently, IU ScholarWorks has added the capability for researchers to store and publish data sets through the IU ScholarWorks service.

More generally the IU Digital Library Program continues to expand its support for digital data collections. Since 2005, the number of digital data collections supported and provided by the DLP has risen from 13 to 34. The total amount of data stored has increased from 14.5 to 32 TB during this time. The Digital Library Program now stores and provides access to more than 1.5 million digital objects.

UITS geographical information systems specialists have collaborated with state and county officials to expand availability of Indiana GIS data. Since 2005, the amount of data stored by IU and accessible for use by the IU research community, Indiana government, and the citizens of

Indiana has increased from approximately 2 TB in 2005 to 18.7 TB today. During calendar 2009, a total of 442,860 accesses of GIS data made available by IU resulted in an aggregate of 11.33 TB of data downloaded with 13,025 distinct hosts served.

CRT Recommendation #7: *Indiana University should continue without pause its substantial investments in high performance computing systems and supercomputers, addressing the diverse needs of IU researchers, clinicians, engineers, and artists. IU should in particular focus on high performance computing for data-intensive scholarship.*

IU acquired its Big Red supercomputer in 2006, originally with 20.4 TFLOPS of computational capability (20.4 trillion mathematical operations per second). Big Red made its debut in 23rd place on the list of the 500 fastest unclassified supercomputers in June of 2006. It was upgraded to 30.7 TFLOPS the following year. Big Red has won acclaim locally and nationally, and has opened the doors to new possibilities in scientific research at Indiana University. We have added computational capability with the Quarry and PolarGrid systems. Most recently, IU has added several modest-sized supercomputers for the FutureGrid project, totaling more than 17 teraflops of computational power with an additional system to be added in 2010. Figure 1 demonstrates, however, how rapidly the field of supercomputing moves; Big Red did not make the list in June of 2010. Staying competitive with leading research institutions requires constant investment. The facilities we are adding through the $15M FutureGrid project will aid IU research ($10M of this funded by the National Science Foundation (NSF)). However, strong investment, for now in additional supercomputer facilities for IU, and eventually perhaps in cloud computing resources, will be required to keep Indiana University at the forefront of discovery and knowledge creation.

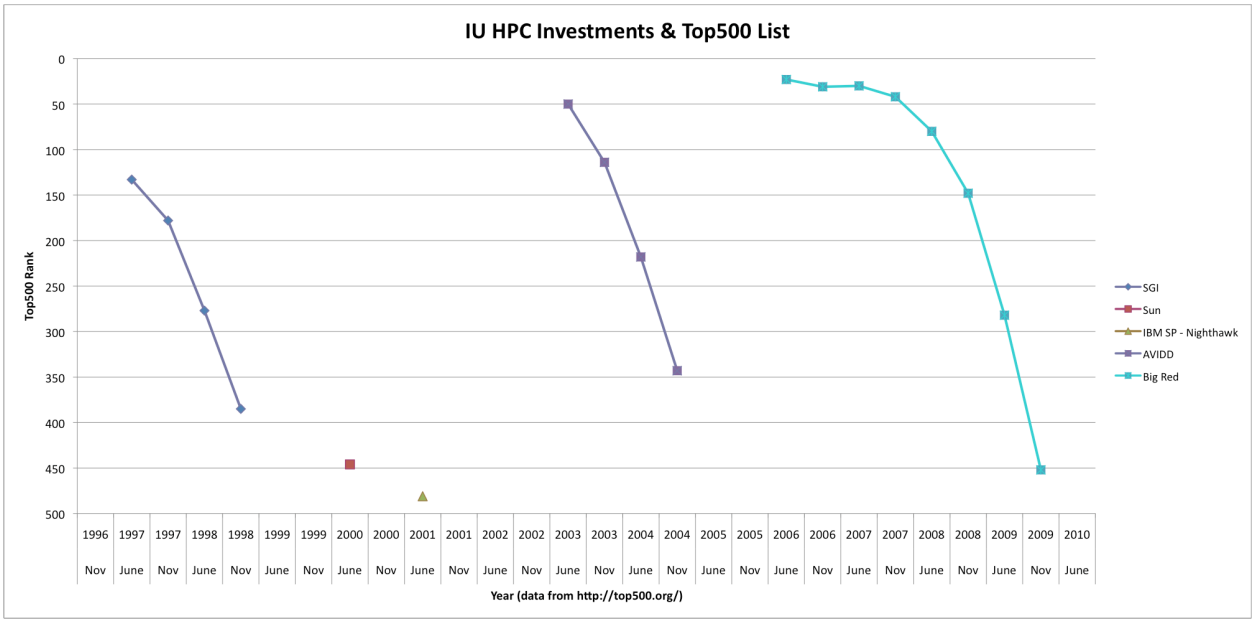

Figure 1. Rank of IU high performance computing systems on the Top500 list (list of the 500 fastest non-classified supercomputers in the world). The highest relative placement ever achieved by an IU system was Big Red's placement on the June 2006 list, at 23rd. Big Red is listed in 452nd place on the November 2009 list. (Data from www.Top500.org.)

CRT Recommendation #8: *Indiana University should continue to invest in a variety of distributed visualization facilities that broadly impact the scholarly and creative efforts of IU*

From 2006-2009, UITS Research Technologies and the Advanced Visualization Lab (AVL) made significant investments in technologies that have provided broad and immediate impact to strategic communities, while also funding forward-looking technology explorations that will result in significant new services and opportunities in 2010 and beyond.

Notable investments for specific communities include new 3D scanning technologies for funded medical research, forensics, sculpture, and artifact digitization; a 3D color printer for artists, ergonomic designers, architects, and physical scientists; a motion capture studio for virtual world designers, game developers, and training simulations (in conjunction with UITS Learning Technologies); interactive information displays for an internationally-traveling exhibit on mapping science and information; a shared Second Life island for conducting group behavior experiments, social networking research, and teaching 3D interaction; and technology installations, telecollaboration facilities, and individual productivity equipment for the recently-established Institute for Digital Arts and Humanities.

Future-looking investments and explorations that will be generating notable results in the year ahead include stereoscopic video, display, and animation technologies; ultra-resolution (defined as greater than four times the resolution of HDTV) displays and content acquisition methods; and low-cost immersive visualization and virtual reality systems. Based on input from the AVL, the new IU Cinema (opening Fall 2010) will include a stereoscopic projection capability, while departments such as Telecommunications, Fine Arts, Music, and New Media are utilizing AVL resources and expertise to develop innovative stereoscopic content. Ultra-resolution display techniques are at the heart of several committed and planned facilities that will serve a broad range of research, presentation, collaboration, and information access needs. Current plans call for a large tiled display in the Innovation Center at IUB (funded in conjunction with IU Research & Technology Corporation), ultra-resolution projection capabilities in the IU Cinema (funded through Cinema initiative), a general purpose tiled display in the public area of the new Cyberinfrastructure Building (planned), as well as several smaller tiled displays funded by specific researcher grants or initiatives. New low-cost immersive visualization systems will continue the trend established with the John-e-Box of providing a technological bridge from high-end facilities like the VR Theater at IUPUI to the labs, classrooms, and studios of researchers, educators, and artists, thereby providing even greater accessibility, utility, and benefit.

CRT Recommendation #9: *Indiana University should foster effective use of cyberinfrastructure through an array of consulting services, complexity-hiding interfaces, and training that will enable scholars to be more innovative and productive (and thus more competitive for grants) through the use of cyberinfrastructure*.

The Pervasive Technology Institute and UITS have demonstrated several "proofs of concept" of the general idea of complexity-hiding interfaces. Indeed, the value of this approach was one of the key points of testimony presented by PTI Executive Director Craig Stewart when he spoke before the House Science and Technology Committee in July of 2008. This area of technology is making the transition from "research" to "deliverable," with some of the prominent research having been done within PTI through projects such as the Open Grid Computing Environments (OGCE) science gateway toolkit and the XBaya graphical workflow composer. One of the most

significant accomplishments to date is the development of the Indiana CTSI HUB – a virtual organization environment enabling collaboration among researchers participating in the IU-led Indiana Clinical and Translational Sciences Institute (http://www.indianactsi.org/). The Indiana CTSI HUB is the portal for this multi-institutional collaboration (IU, Purdue, and Notre Dame) and supports researcher and resource discovery, support of translational research workflows, and tools for collaboration among researchers, industry, and the public. The Indiana CTSI HUB is notable as the first CTSA informatics platform to support Shibboleth-based InCommon logins that allows trusted information sharing among translational research nationally. This particular interface is based on the HUBzero software initially developed by Purdue University. IU is a charter member with Purdue of the HUBzero consortium, which aims to sustain and improve the software that powers this and other complexity-hiding interfaces. Ongoing research within both the PTI's Digital Science Center and Data to Insight Center continues to push the boundaries of scientific computing interfaces, including work in immersive workflow environments and workflow emulation.

CRT Recommendation # 10. *Indiana University should continue to lead and participate in leveraged efforts to develop, deploy, and make use of cyberinfrastructure for the State of Indiana, at national, and international levels*.

IU is becoming increasingly successful and increasingly well recognized as a national leader in development and deployment of cyberinfrastructure. Key accomplishments include:

- Cybersecurity. The Center for Applied Cybersecurity Research (CACR). The CACR provides national leadership in security and privacy policy, and Dr. Fred Cate, the director, frequently testifies before Congress on privacy and technology. The CACR is also engaged in research on netflow tracking of spam and other attacks, on home healthcare devices, and on new models for securing data flows across the Internet.
- Use of Big Red in support of economic development in the state of Indiana. Indiana University, Purdue University, and IBM Inc. have collaborated through the Indiana Initiative for Economic Development (http://rtinfo.indiana.edu/IIED). This initiative has enabled companies from startups to some of Indiana's largest employers to achieve innovation and business advantage through use of a portion of Big Red provided to the state for this purpose, by IBM Inc.
- Creation of Polar Grid – a nationally used cyberinfrastructure for measuring polar ice sheets (supported by NSF). As a nation we need clear, accurate data as to the current status and rate of change of polar ice sheets. IU's Polar Grid project provides the cyberinfrastructure to do that – computing equipment and IT experts sent into the field in the Arctic and Antarctic, and a major supercomputer system in Indiana to better analyze data collected in subzero temperatures at our globe's poles.
- The Global Research Network Operations Center (GRNOC) earned its "Global" name through its record in international networking, beginning with the creation of the TransPAC and TransPAC2 networks interconnecting the research cyberinfrastructures of Asia with those of the US. TransPAC2, a grant awarded to IU in 2005, has allowed the GRNOC to take a leadership role in the establishment of connectivity between scientists and researchers in Pakistan and their counterparts in the global scientific community. The GRNOC has also recently been awarded an NSF grant to support development of

operational and measurement tools for the GENI project (http://gmoc.grnoc.iu.edu/), to be used by researchers as they create the next generation of networks for advanced research.

- The GRNOC sits at the heart of US advanced networking as the operations and network engineering arm of Internet2 and National Lambda Rail, the two pre-eminent advanced national backbone networks supporting the most important research and education efforts of the nation's higher education, government, and primary education communities. The GRNOC supports Indiana's own statewide optical network for higher education institutions, and over 15 external organizations have contracted with the GRNOC to support networks across the country. State education networks, large regional optical networks, and optical exchange points are some of the projects now supported by the GRNOC.
- Participation in operations of the TeraGrid and the Open Science Grid, the two leading unclassified grid computing systems in the US. These two major grid projects are the state of the art in grid computing, and IU plays a leadership role in both.
- The 2009 $10.1M grant from the National Science Foundation to Indiana University is a new high-water mark in IU achievement in research cyberinfrastructure. FutureGrid, led by PI Geoffrey C. Fox and supported by the Research Technologies Division of UITS and PTI, will help define the future of grid and cloud computing. This innovative, national testbed puts IU at the forefront of research and development in cyberinfrastructure. This project also assures IU a place in the NSF-funded TeraGrid until 2013 – IU being one of just 5 institutions with a place secured within the TeraGrid past 2011. IU's involvement in the TeraGrid has tremendous value to IU in terms of grant competitiveness generally and value to IU researchers who have need for the most advanced cyberinfrastructure in the US.
- Open Science Grid (OSG). IU operates the Grid Operations Center (GOC) for the Open Science Grid, which includes more different individual computers than any other computing grid on earth. In total, the Open Science Grid includes thousands of computers worldwide, but in particular many advanced computer systems in the US dedicated to analysis of data from the Large Hadron Collider (LHC). The IU OSG group has developed a new tool, called MyOSG, which is a web portal that consolidates and presents information to create custom user views from multiple grid data sources. IU originally developed MyOSG to help monitor the many distributed high performance computing systems that make up the OSG. MyOSG technology proved to be successful and easy to use, and was recently adopted by Enabling Grids for E-sciencE (EGEE), a European grid infrastructure that supports the LHC.