Usage of UITS advanced research cyberinfrastructure for 2011

Matthew R. Link

Indiana University

PTI Technical Report PTI-TR12-009

10 April 2012

Citation:

Link, M.R. "Usage of UITS advanced research cyberinfrastructure for 2011." Indiana University, Bloomington, IN. PTI Technical Report PTI-TR12-009, April 2012. Available from: http://hdl.handle.net/2022/14352

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

1. Cyberinfrastructure origins and outcomes

IU has a proud tradition in open access to its research computing and cyberinfrastructure (CI) facilities, going back to the precedents set by Marshall Wrubel (appointed the first permanent director of the IU Research Computing Center in 1955). Starting in 1997 President Myles Brand and then-Vice President Michael McRobbie initiated a tremendous acceleration in growth of IU's cyberinfrastructure facilities through developing and then executing the first Indiana University Information Technology Strategic Plan. Through a decade and a half of purposeful execution of excellent strategies in support for research and scholarly activities generally, University Information Technology Services (UITS) has provided exceptional support to a group of researchers. This includes usage from disciplines that are among the traditional users of high performance computing – physics, chemistry, and astronomy, as well as emerging areas of application of HPC including biology, business, and the arts.

UITS, particularly the Research Technologies (RT) Division, is supporting research – particularly federally funded research – for a wider variety of disciplines than ever. RT is now being invited to participate in major national cyberinfrastructure projects on the basis of our expertise, largely without match or with minimal match funding from IU. RT is aiding IU's success in research – particularly the Pervasive Technology Institute (PTI), and recruitment and retention of faculty and staff in ways that are significantly aiding IU's missions of research, education, and engagement in the state.

More than 14% of the Bloomington campus research community and more than 10% of the IUPUI community currently use IU's advanced cyberinfrastructure. Comparable figures are hard to come by, but our peers and competitor institutions informally report usage figures in single digit percentages. Use of IUs advanced research cyberinfrastructure is so widespread that it is no longer possible to track effectively the number of scholarly publications produced by IU that make use of these resources – the number of such publications is too high.

We can effectively track some of the key financial metrics relative to IU's advanced cyberinfrastructure. Aggregate research awards to IU in support of information technology and informatics research, including PTI and its predecessor, Pervasive Technology Labs, now total $173,016,092 since 1999. Considering very narrowly grant awards that depended very heavily on the IU Data Center (e.g. FutureGrid) or grants led by Research Technologies, a total of $31,342,933 in NSF and NIH funding has been awarded competitively to IU; of this $6,528,078 was facilities & administration funds.

The foundation around which all of this success has been built is IU's strategic and ongoing investment in advanced cyberinfrastructure systems, deployment, and support by world-class staff with both technical and discipline-facing expertise. OVPIT has recently funded major upgrades to archival data storage systems, and funding for visualization environments is pending. These areas of cyberinfrastructure managed by RT are presently in (or on their way to being) well matched to IU's needs for now and the next few years.

2. Background and supporting information

In 1997, IU acquired a supercomputer with assistance of NSF funds for the first time. The IU Information Technology Strategic Plan of 1998[1] put significant emphasis on research computing, storage, and visualization facilities general – those things we now call cyberinfrastructure. A string of activities and intellectual successes for IU resulted from the execution of this strategic plan. In 2001 IU upgraded its IBM SP supercomputer to just over 1 TFLOPS – the first university owned supercomputer to exceed the 1 TFLOPS mark. Other projects, including AVIDD, the Data Capacitor, and IU's distributed HPSS installation built up IU's research cyberinfrastructure. Critical high points in this history of IU

[1] http://ovpit.iu.edu/strategic/

supercomputing were the deployment of Big Red in 2006 and its doubling in capacity a year later. Big Red made its debut in the Top500 list in June of 2006 as the 23rd fastest (unclassified) supercomputer in the world, and the fastest in the western hemisphere. Another high point came in 2008, when IU won the Supercomputing Conference Bandwidth Challenge supporting rapid data transport and distributed scientific workflows with the Data Capacitor.

IU has supported an ever-increasing set of research and researchers with its advanced computing environment, as evidenced in figures below. Figure 1 shows the aggregate computational capacity of IU's research cyberinfrastructure.

Figure 2 shows usage in terms of TFLOPS-hours of IU's primary production HPC systems – Big Red and Quarry – from FY 06/07 through FY 10/11. Figure 3 shows usage of IU's parallel file systems over the same time period.

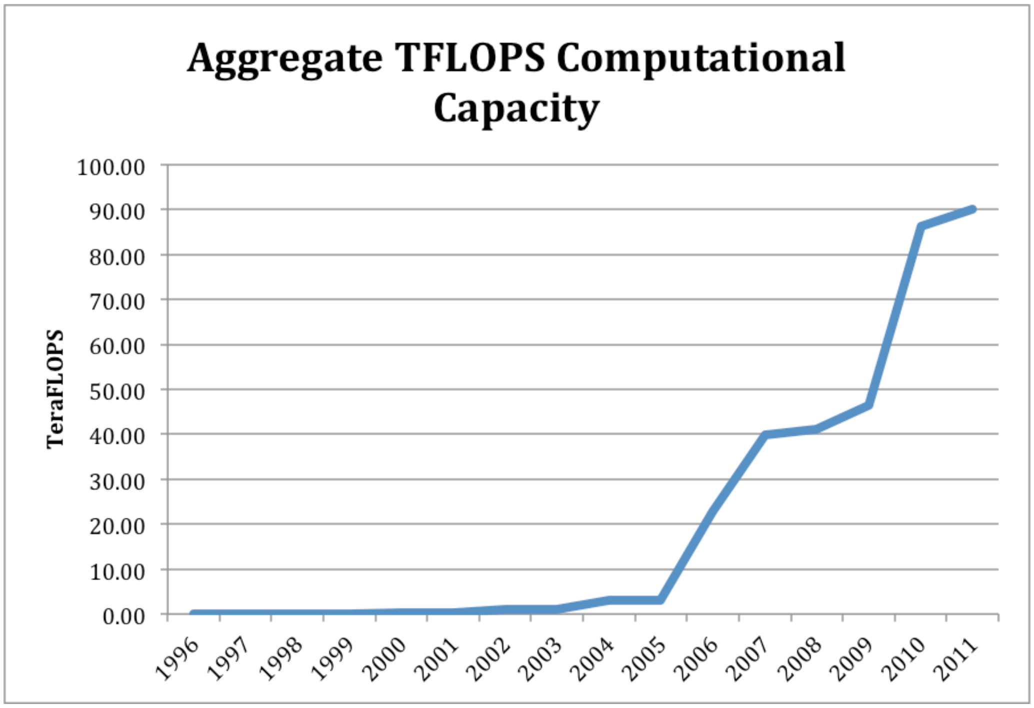

Figure 1. Growth in aggregate TFLOPS computational capacity since 1996.

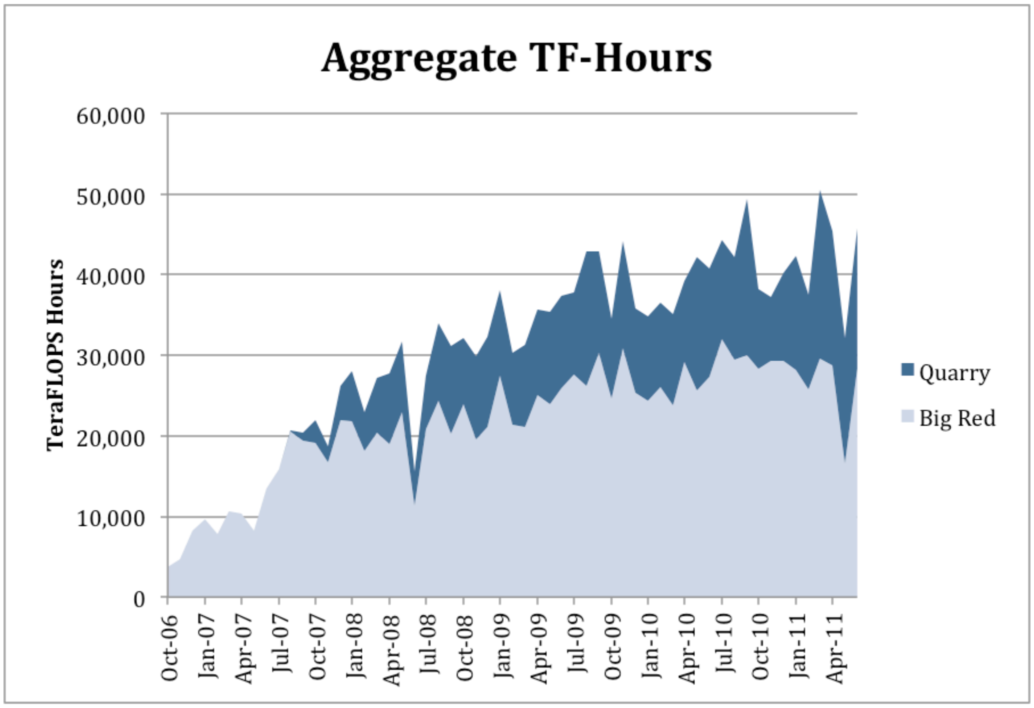

Figure 2. TFLOPS-hours of computing time used on Big Red and Quarry high performance computing systems at IU from FY 06/07-FY 10/11.

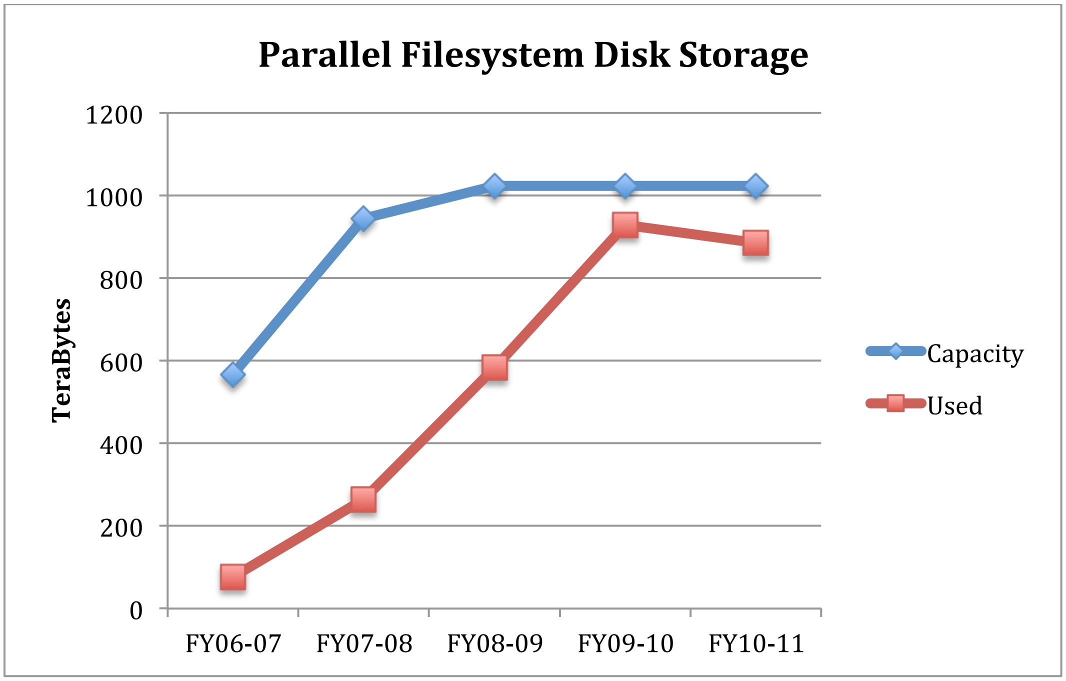

Figure 3. Use of parallel file systems from FY 06/07 to FY 10/11.

Figure 3 shows a slight decline in usage of parallel file systems between FY 09/10 and 10/11. This does not reflect a decrease in demand. Research Technologies worked actively to move data from parallel data file systems to archive tape in response to the Data Capacitor being dangerously close to its total physical capacity. On one occasion data accumulating as part of a workflow almost pushed the Data Capacitor beyond its total capacity (and in fact would have done so without intervention of Data Capacitor systems administrators). Disk space insufficient to meet demand has thus resulted in movement of data to tape even when the uses of such data are best supported by disk.

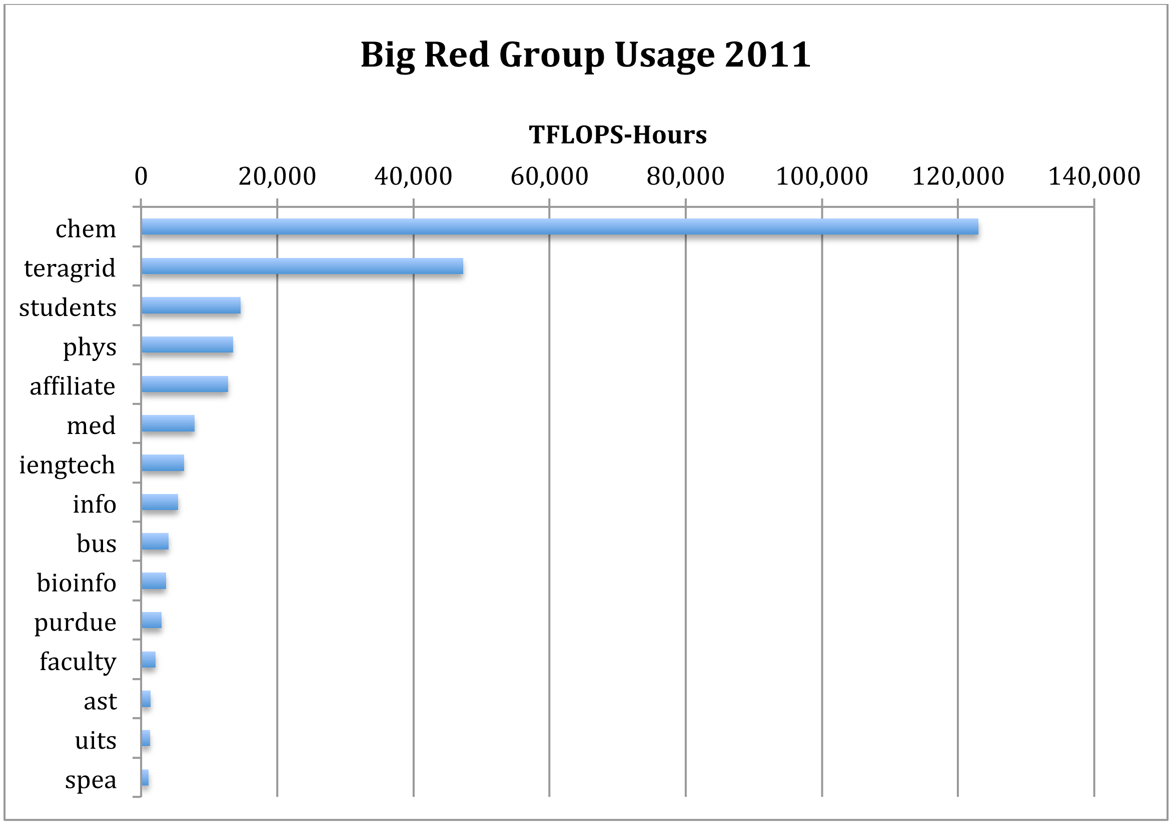

Figure 4. Usage of Big Red by group for CY2011 for top groups. Legend: chem = Chemistry; teragrid = TeraGrid; students = students whose primary association is listed generically in IT Accounts; phys = Physics; affiliate = affiliate accounts; med = School of Medicine or Department of Medicine; iengtech = Indianapolis Engineering & Technology; info = School of Informatics and Computing; bus = Kelley School of Business; bioinfo = Bioinformatics; purdue = Purdue affiliates; faculty = faculty whose primary association is listed generically in IT Accounts; ast = Astronomy; uits = UITS/OVPIT researchers, informatics and IT research; spea = School of Public & Environmental Affairs.

As regards Figure 4, it is important to note that IU's 2005 proposal to become a Resource Partner in the TeraGrid committed 30% of Big Red's initial 20.48 TFLOPS capacity for use by the national research community. In the end, we delivered 33% – some of which was actually used by IU faculty.

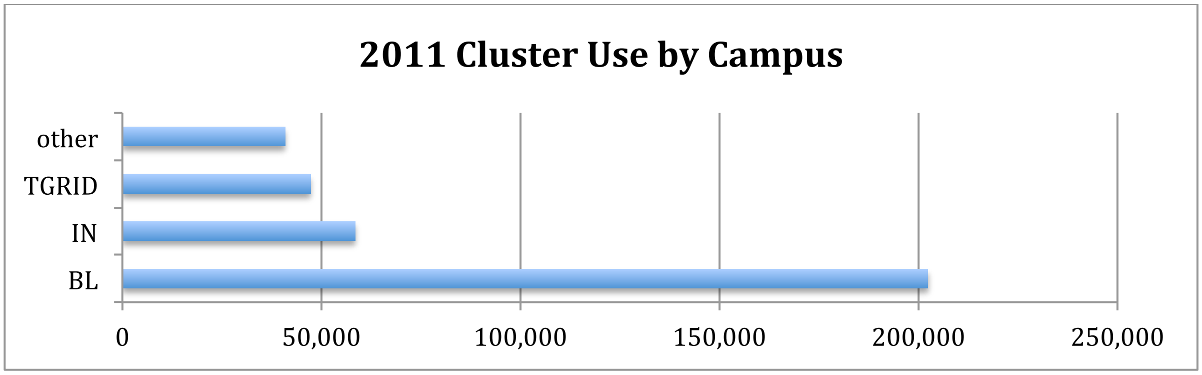

Figure 5: Cluster usage in TFLOPS-hours for CY2011 by campus (Big Red, Quarry, Mason).

Figure 5 shows the usage by campus of all central research clusters during calendar year 2011. Campus affiliation data were obtained from the IU Account Management System (AMS), and the "other" category includes the indeterminate cases where it's not possible to tell with what campus a user is affiliated due to multiple affiliations.

User Rank (largest user of Big Red first)	Department	TFLOPS-Hours	% of total system utilization
1	IUB-CHEM	35,265	12.88%
2	IUB-CHEM	32,519	11.87%
3	IUB-CHEM	23,900	8.73%
4	TERAGRID	16,891	6.17%
5	UA-VPIT	10,314	3.77%
6	IUPUI-BIOM	10,096	3.69%
7	IUPUI-CHEM	9,888	3.61%
8	TERAGRID	8,292	3.03%
9	TERAGRID	7,380	2.69%
10	TERAGRID	6,073	2.22%
11	IUPUI-ENGT	5,618	2.05%
12	IUB-CHEM	4,912	1.79%
13	IUPUI-CHEM	4,682	1.71%
14	IUB-CHEM	4,170	1.52%
15	IUB-PHYS	4,070	1.49%
16	IUB-BUS	4,028	1.47%
17	IUPUI-BIOM	3,384	1.24%
18	IUB-CHEM	3,337	1.22%
19	IUB-CEEM	3,264	1.19%
20	IUB-PHYS	3,010	1.10%
21	IUPUI-PHTX	2,747	1.00%
22	IUPUI-ENGT	2,468	0.90%
23	IUB-CHEM	2,432	0.89%
24	IUB-CHEM	2,372	0.87%
25	IUB-CHEM	2,149	0.78%

Table 1: Characteristics (campus, department, TFLOPS-hours used, % of total system hours) of the 25 individual top users of Big Red during CY2011. Legend: BL-BUS = IU Bloomington, Kelley School of Business; IUB-CEEM = Center for Exploration of Energy and Matter; IUB-Chem = Chemistry at IU Bloomington; IUB-Physics = IU Bloomington, Physics dept.; IUPUI-BIOM = IUPUI, Biomedical research (bioinformatics related); IUPUI-ENGT = IUPUI, Engineering & Technology; IUPUI-PHTX = Pharmacology and Toxicology (School of Medicine); TERAGRID = Non-IU users, accessing IU systems via allocations

through the NSF-funded TeraGrid project and associated grant awards to IU; UA-VPIT = all OVPIT affiliated accounts, IT and informatics research.

Table 1 above provides information about the types of disciplines that have been the top users of Big Red (Big Red is much larger than Quarry; consequently, many of the large parallel jobs run on Big Red. This table summarizes the top computational users of IU's HPC servers overall).

Table 2 below shows the usage of IU's HPC facilities among the rank and file of IU faculty, staff, and graduate student researchers. In terms of usage within the general university community we believe that is without peer among other institutions of higher education.

	IUB			IUPUI		
	Average	**Satisfaction**	**Usage**	**Average**	**Satisfaction**	**Usage**
Central research and high performance computers (Big Red, Quarry, and RDC clusters) [F, Staff, G]	4.16 ± .08	92.9 ± 2.2%	14.6%	4.19 ± .07	95.0 ± 1.9%	10.9%
Center for Statistical and Mathematical Computing (Stat/Math Center, statmath@indiana.edu, 855-4724/278-4740) [All]	4.08 ± .06	94.0 ± 1.7%	20.1%	4.02 ± .08	93.4 ± 2.1%	8.4%
Massive Data Storage Service (MDSS/HPSS) [F, Staff, G]	4.16 ± .07	93.5 ± 2.1%	16.2%	4.09 ± .08	96.0 ± 1.7%	10.1%
Advanced Visualization Laboratory (AVL); www.avl.iu.edu [F, Staff, G]	4.32 ± .07	97.8 ± 1.3%	7.8%	4.17 ± .07	96.7 ± 1.6%	6.0%
Support for software applications using IU's high performance computing resources [F, Staff, G]	4.11 ± .08	93.3 ± 2.1%	15.2%	4.16 ± .07	96.6 ± 1.6%	11.7%
Support for life sciences research (Bioinformatics and the Center for Computational Cytomics (IUB); Biomedical Applications and Advanced IT Core (IUPUI)) [F, Staff, G]	3.91 ± .08	91.5 ± 2.3%	8.3%	4.22 ± .08	96.0 ± 1.7%	8.7%
Overall satisfaction	4.17 ± .06	95.5 ± 1.5%	50.0%	4.30 ± .05	98.5 ± .9%	51.6%

Table 2. Summary of usage and satisfaction of IU research cyberinfrastructure from the 2011 UITS User Satisfaction Survey (http://www.indiana.edu/~uitssur/).

A summary of key financial metrics relative to IU's advanced cyberinfrastructure is presented in Table 3.

Project	IU investments	NSF funding for hardware	NSF/NIH direct costs for services, software, and support	Facilities and Administration funds to IU
Big Red	$9,000,000	$0	$0	$0
Quarry	$905,370	$0	$0	$0
Data Capacitor and other spinning disk storage	$758,822	$0	$0	$0
Archival research storage (HPSS)	$2,939,457	$0	$0	$0
TeraGrid and XSEDE RP awards and management subawards	$36,554	$81,995	$7,853,715	$4,105,487
FutureGrid	$5,870,781	$1,250,766	$3,877,798	$822,275
Polar research	$863,125	$1,583,224	$3,020,297	$694,878
Data Capacitor	$0	$1,720,000	$1,720,000	$0
Open Science Grid	$174,494	$5,100	$1,527,847	$782,085
Life sciences software	$485,000	$0	$1,869,083	$632,208
CI software studies	$0	$0	$ 305,030	$123,353
Totals	$21,033,603	$4,641,085	$20,173,770	$6,528,078

Table 3. Summary of key financial metrics related to IU research cyberinfrastructure from July 1 2005 to August 2011 – IU investments in major projects and NSF and NIH funding to IU (and spent by IU – not disbursed via subcontracts).

A recent paper[2] presented a statistical analysis of publically available research data and information on ownership of high performance computers (HPC) in the list of the 500 fastest supercomputers in the world[3]. This study led to the following conclusions: “Appearance on the Top 500 list is associated with a contemporaneous increase in NSF [National Science Foundation] funding levels as well as a contemporaneous increase in the number of publications. … consistent investments in HPC at even modest levels are strongly correlated to research competitiveness.” A positive relationship between investment in high performance computers and aggregate funding from the National Science Foundation was also noted.

Recent activities led by the Research Technologies Division show how far this part of UITS has come since 1997:

- IU’s subcontract as part of the 5-year XSEDE program is $4,975,000 (with potential for revision upward) with minimal match from IU). IU leads XSEDE in the area of campus bridging.
- IU will have a subcontract on the as-yet-unannounced award to TACC for the next Track II system – called Stampede.
- RT was recently granted an award through the NSF EAGER mechanism for $296,637 to do a study of software sustainability success cases.
- RT is coordinating a major new initiative to provide “cluster as a service” from Penguin Computing Inc., located physically in IU’s Data Center, to provide flexible “above campus” services and obtain in return cycles at no cost to use in support of the IU community.
- Grant awards led by RT are providing a larger amount of funding to other divisions of UITS than ever before.

[2] Apon, A., S. Ahalt, V. Dantuluri, C. Gurdgiev, M. Limayem, L. Ngo. 2010. High Performance Computing Instrumentation and Research Productivity in U.S. Universities. Journal of Information Technology Impact 10: 87-98

[3] www.top500.org

The cyberinfrastructure and services provided by the Research Technologies Division has been a foundation on which much of the success of the Pervasive Technology Institute has been built (the other foundational aspect being of course the excellence of the faculty research leaders). IU is aiding recruitment of faculty, and retaining staff, in ways that are significantly aiding IU's missions of research, education, and engagement in the state.

3. Services for the College of Arts and Sciences

The College continues to use an extremely large share of UITS research cyberinfrastructure resources delivered by RT. During calendar year 2011 the College used 4,980,439 node hours of supercomputing time on Big Red and Quarry. This was 43% of the time utilized on those systems. IU's internal cost for this usage was $896,479 (based on IU's Activity Based Costing figures for FY11). There is no good basis for cost comparisons on the open market because these services cannot be purchased on the open market. The closet analog available for cost comparison suggests that this amount of resource would have cost between $2.5M and $5M.

The College has 162 terabytes (TB) of data stored on spinning disk in the Data Capacitor and Data Capacitor WAN systems, amounting to 21% of the available capacity. The cost of storing this much data on disk, based on IU internal costs for 2011, is $161,877. Amazon S3 cloud storage would cost at least $199,065/year for the equivalent amount of storage, not counting the cost of moving data in an out of such storage systems. Based on usage patterns, data movement costs would also have been also in the hundreds of thousands of dollars.

The College had 1.7 petabytes (PB) of data stored in the UITS Scholarly Data Archive (SDA) as of the end of 2011. SDA is a tape storage system that replicates data between Bloomington and Indianapolis for disaster preparedness, so the total amount of tape used is 3.4 PB. The College represents 44% of the data stored in SDA. The College used a similar percentage of storage in the Research File System, which provides spinning disk storage to the desktop. The internal cost to IU for tape storage in FY12 is $572/TB, or $972,000 for the College's data. Amazon S3 cloud storage would cost at least $1,956,000/yr for the equivalent amount of storage. Again, cost of moving data in and out of Amazon services would be in the additional hundreds of thousands of dollars.

UITS has supported the implementation of a number of small-scale to room-scale visualization systems throughout the university, greatly expanding access to large-scale visualization facilities. UITS has in particular supported visualization related to sciences and the fine arts (including supporting installations by Associate Professor Margaret Dolinsky and 3D printing activities of Assistant Professor Nicole Jacquard).

UITS and PTI have collaborated with the Astronomy Department and the WIYN (Wisconsin Indiana Yale NAOA) consortium to develop new tools for managing and analyzing data from the One Degree Imager (ODI) telescope being developed by WIYN. UITS and PTI are developing a software pipeline to process raw data and create "science usable" images. As part of this agreement, matching contributions of use of UITS resources are being counted toward the College's financial commitments to WIYN, and these contributions will be in excess of $100,000.

- The ODI is a 1-gigapixel camera being built by WIYN, Inc. to be installed in an existing 3.5m ground-based telescope located at Kitt Peak Mountain, AZ. ODI will be one of the best ground-based telescopes in existence when it begins operation. WIYN is a 501(c)(3) organization supported by of the University of Wisconsin, Indiana University (in particular the IU Department of Astronomy within the College), Yale University, and NOAO.
- ODI's larger-than-current-norms data sizes lead to the need for a science gateway to enable astronomers to reach the full scientific potential of ODI. Since January 2010, RT has teamed up with the College and other WIYN partners to propose and later to design a science gateway called

ODI-Pipeline, Portal, and Archive (ODI-PPA); the design includes the use of the IU Scholarly Data Archive, the Data Capacitor, the Quarry cluster, and open source software developed at IU. RT's management personnel and hardware contributions have significantly reduced the financial expectations for the College (toward their partnership within WIYN), while the project has enabled RT to broaden its interdisciplinary collaborations to include the field of Astronomy.

- An external Critical Design Review panel recently reviewed the project favorably. The panel added (emphasis added): "This design is also a forerunner of what is likely to become a ubiquitous approach to data processing in astronomy ... In short, *the success of ODI-PPA will likely pave the way for similar systems at other astronomical observatories*." The ODI-PPA science gateway will be developed in 2012-13; the partnership between UITS Research Technologies, the College (IU Astronomy), and WIYN is expected to continue to be a tremendous success and is expected to be a major boost toward the College/IU Astronomy's plans as well as potential RT-Astronomy partnerships in the future.

The Pervasive Technology Institute grew and was reorganized during the last year. PTI consists of two types of centers: Research Centers, which are traditional faculty-led research groups; and Service and Cyberinfrastructure Centers. Managing Director Beth Plale, Professor in the School of Informatics and Computing (SOIC) coordinates activities of the Research Centers. Executive Director Craig Stewart coordinates activities of the Service & Cyberinfrastructure Centers. Stewart and Plale are jointly responsible for the overall success of PTI. Some critical successes of PTI relevant to the College, and particularly relevant to the new academic directions charter, are the following:

- The Data to Insight Center, led by Professor Plale, has led the creation of the HathiTrust Research Center, which supports research on digitized texts that are part of the HathiTrust repository (over 10 million digitized volumes and counting – see http://www.hathitrust.org/about for more information).
- The Data to Insight Center (D2I) is one of the leading partners in an NSF-funded project called "Sustainable Environments – Actionable Data" that is focused on collecting and providing access to a variety of data related to sustainability science, which will be of use to many IU scholars in sociology, political science, economics, and SPEA).
- Network expert Martin Swany has joined the School of Informatics and Computing and D2I. His expertise is in movement and management of large data sets, and he is developing new software tools that will aid researchers in the College who need to access very large data sets.
- PTI has added a new center – the Center for Research in Extreme Scale Technologies, led by SOIC Professor Andrew Lumsdaine, with recently recruited SOIC Professor and OVPIT Distinguished Scientist Thomas Sterling. (Professor Sterling is widely known as the "father of Beowulf clusters.") Technology being developed by CREST will be particularly valuable in programming the largest supercomputers in the world to address challenges in genomics, physics, and astronomy.
- The Digital Science Center, led by SOIC Professor and OVPIT Distinguished Scientist Geoffrey Fox, is developing new cloud-based technologies of particular importance to bioinformatics. These are being implemented and tested on FutureGrid, a national cloud and grid testbed led by Professor Fox.
- UITS and PTI, in collaboration with the Department of Biology, received $1.5M in funding from the NSF to create the National Center for Genome Analysis Support (NCGAS). NCGAS represents the first formal organizational relationship between the College and PTI, and is formed as a new Service and Cyberinfrastructure Center affiliated with PTI. (The other affiliated Service and Cyberinfrastructure Center is the Research Technologies Division of UITS). NCGAS will aid

researchers nationally and locally with deployment and use of new applications in genome assembly and metagenomics. (PI Craig Stewart leads NCGAS, with William Barnett of UITS, Michael Lynch and Matthew Hahn of Biology, and Geoffrey Fox of SOIC and PTI as co-PIs). As part of this initiative, UITS deployed the Mason cluster, a large memory (512GB per node) cluster designed specifically to meet needs of IU biologists. IU has also established a partnership with Penguin Computing, Inc. to establish an "on demand" service located at IU available to researchers nationally and locally – providing secure options to meet needs that can be characterized as "I have money and I want to compute now."

- Several subunits of PTI have collaborated to create templates and guidance for IU researchers to use in preparing data management plans as now required by the NSF as part of proposals to the NSF. These materials emphasize use of IUScholarWorks as a way to store data for dissemination, and use of the open software statistical package R, as ways to ensure open access to data and the ability to repeat and extend analysis of data done as part of NSF-funded research.

4. Services for the IU School of Medicine

Research Technologies works on several research support projects for faculty from the IU School of Medicine and Indiana Clinical and Translational Sciences Institute (CTSI); Regenstrief Institute; the IU Simon Cancer Center; Chemistry and Chemical Biology; the Center for Computational Biology and Bioinformatics; Indiana University Northwest; and others. The most active area of service development and delivery is the Indiana CTSI HUB (www.indianactsi.org), which RT provides as a statewide online translational research portal to the Indiana CTSI. As part of this program, Dr. William Barnett's CTSI ARRA supplement proposal (Anantha Shekhar, PI) to build new HUB tools was funded, allowing continued development of the CTSI HUB. As part of this award, the RT Advanced IT Core (AITC) launched four new services:

- i2iconnect (www.i2iconnect.org): An online service that provides licensing matchmaking between inventors and technology transfer officers and the healthcare industry. Notably, i2iconnect was a Techpoint Mira Awards finalist for 2011 in the Innovation of the Year category.
- REDCap (http://redcap.uits.iu.edu): A clinical data collection and management platform.
- AlfrescoShare (http://alfresco.uits.iu.edu): A multi-institutional secure file sharing service.
- INResearch.org (http://inresearch.org): A clinical research volunteer recruitment portal for the Indiana CTSI

Additionally, through subsequent Clinical and Translational Science Award administrative awards led by Dr. Barnett, RT provided two additional translational research services

- CTSA2Community (ctsa2community.org): A best practices repository for community engaged research
- Multi-team online workflows: Using technologies developed for the Cancer Care Engineering HUB (cceHUB.org), RT partnered with cceHUB to develop web-based workflows that support multi-team investigations and provide tools for data analysis and management that accelerate collaborative research.

Dr. Barnett continues as the PI of the Informatics Core for the Collaborative Initiative on Fetal Alcohol Spectrum Disorders (CIFASD) led by Dr. Ed Riley (San Diego State University). The AITC also acts as the Informatics Core for the National Gene Vector Biorepository led by Dr. Ken Cornetta (Medical and Molecular Genetics, IUSM). Ganesh Shankar, manager of the AITC, continues as Informatics lead for the Neuro-AIDS research program (Yiannoutsis, PI) and the caBIG integration of OnCore and Regenstrief data. The AITC completed the Life Stress project, for Dr. Delunas at IU Northwest, which allows participants to share the impact of stressful events on their health, and receive information for coping with stress.

Other RT projects that benefit the IU School of Medicine include:

- Identification of Biomarkers for Predilection to Drug Addiction. RT deployed a data transport and analysis platform for Dr. Hulvershorn and collaborators who are using functional MRI techniques to identify biomarkers for drug addiction in adolescents.
- VIVO Integration. For the Indiana CTSI, and as part of the VIVO initiative to develop a national semantic web architecture of faculty profile information to support research collaboration and administration, RT has established the Indiana CTSI HUB as a node on the VIVO experimental network.
- VIVO Mini-Grant. Dr. Barnett has received a mini-grant from the VIVO project to create middleware for the HUBzero Web 2.0 environment that will allow HUB software applications to access linked open data (semantically structured open data sets that allow intelligent querying among different, diverse data repositories). This is foundational cyberinfrastructure for integrating diverse data in support of research. The RT team is also developing some pilot applications that will access faculty profile information for collaborative science applications.