Extensible Terascale Facility (ETF): Indiana-Purdue Grid (IP-Grid)

Final Report

NSF Award ID: ACI-0338618
Project Dates: 10/1/2003 to 9/30/2005
Principal Investigator: McRobbie, Michael
Co-Investigators: Fox, Geoffrey C., Gannon, Dennis B., Palakal, Matthew J., Voss, Brian D.
Organization: Indiana University

Table of Contents

Final Report for Period: 10/2003 - 09/2005
Submitted on: 01/09/2006
Principal Investigator: McRobbie, Michael A.
Award ID: 0338618
Organization: Indiana University
Title:
Extensible Terascale Facility (ETF): Indiana-Purdue Grid (IP-grid)

Project Participants

Senior Personnel

Name: McRobbie, Michael
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Gannon, Dennis
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Palakal, Mathew
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Fox, Geoffrey
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Voss, Brian
Worked for more than 160 Hours: Yes
Contribution to Project:
Brian D. Voss left Indiana University in March, 2005. Up until his departure, Brian worked extensively on TeraGrid related activities.

Post-doc

Graduate Student

Undergraduate Student

Technician, Programmer

Other Participant

Research Experience for Undergraduates

Organizational Partners

UNIVERSITY OF ILLINOIS, NCSA

Oak Ridge National Laboratory

Pittsburgh Supercomputing Center

Purdue University

University of Texas at Austin

University of California-San Diego

University of Chicago

Other Collaborators or Contacts

Charles Little at the University of Kansas provided specifications and data for embryology data.

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)
See attachment for description of major research and education activities.

Findings:
See attachment for description of major findings from research and education activities.

Training and Development:
See attachment for research and teaching skills and experience this project helped provide.

Outreach Activities:
See attachment for outreach activites provided to the public to increase understanding and participation in science and technology.

Journal Publications

C. J. Horowitz, M.A. Perez-Garcia, D.K. Berry, and J. Piekarewicz, "Dynamical response of the nuclear "pasta" in neutron star crusts", Phys. Rev. C, p. 72, vol. , (2005). Published

Anurag Shankar, "Grid computing - getting connected", Linux for you, p. 22, vol. , (2004). Published

Craig A. Stewart, "Bioinformatics: transforming biomedical research and medical care", CACM, p. 30, vol. 47, (2004). Published

Mark Ellisman, Michael Brady, David Hart, Fang-Pang Lin, Matthias S. Muller, Larry Smarr, "The emerging role of biogrids", CACM, p. 52, vol. 47, (2004). Published

Books or Other One-time Publications

Maguitman, A., A. Rechsteiner, K. Verspoor, C. Strauss, L. Rocha, "Large-Scale Testing of Bibliome Informatics Using Pfam Protein Families", (2006). Proceedings of the PSB 2006, Published
Bibliography: N/A

Maguitman, A., F. Menczer, H. Roinestad, A. Vespignani, "Algorithmic Detection of Semantic Similarity", (2005). Proceedings of the WWW 2005, Published
Bibliography: N/A

McRobbie, M.A., and C.A. Stewart, "Pervasive Technology Labs 66 month report for Lilly Endowment from Indiana University Bloomington, IN", (2005). Report, Published
Bibliography: N/A

Stewart, C.A. R. Keller, R. Repasky, M. Hess, D. Hart, M. Mueller, R. Sheppard, U. Woessner, M. Aumueller, Huian Li, D.K. Berry, J. Colbourne, "A global grid for analysis of arthropod evolution", (2004). Proceedings of Fifth IEEE/ACM International Workshop on grid computing 328-337, Published
Editor(s): R. Buyya. Pittsburgh, PA..
Bibliography: N/A

Stewart, C.A., R. Repasky, A. Arenson, "Open source tools for computational biology", (2005). Proceedings/tutorial at SC2004, Pittsburgh, PA., Published
Bibliography: N/A

Stewart, C.A., "Scientific Data Management", (2005). Proceedings/tutorial at PittCon 2005, Orlando, FL., Published
Bibliography: N/A

Stewart, C.A., "Scientific Data Management", (2004). Proceedings/tutorial at PittCon 2004, Chicago, IL., Published
Bibliography: N/A

Stewart, C.A., "Computational Biology", (2003). Proceedings/tutorial presented at SC2003, Phoenix, AZ., Published
Bibliography: N/A

Wernert, E.A., and G. Bertoline, "Proceedings of the 2005 I-light Symposium, Indiana University Purdue University Indianapolis. Indianapolis, IN.", (2005). Proceedings, Published
Bibliography: N/A

Wernert, Eric, Mike Boyles, John N. Huffman, Jeff Rogers, John C. Huffman, and Craig Stewart, "The John-e-Box: Fostering Innovation, Inclusion and Collaboration through Accessible Advanced Visualization", (2005). Proceedings of the Richard Tapia Diversity in Computing Conference 2005 Albuque, Published
Bibliography: N/A

Web/Internet Site

URL(s):

http://iu.teragrid.org/ETF

Description:

This site is a compilation of multiple, various ETF sites at Indiana University which relate either directly, or indirectly to the NSF ETF TeraGrid award.

Other Specific Products

Product Type:

Data or databases

Product Description:

FlyBase - a database of genetic and molecular data for Drosophila. Includes data on all species from the family Drosophilidae; the primary species represented is Drosophila melanogaster

Sharing Information:

Database is available on TeraGrid to all users as a resource.

Product Type:

Data or databases

Product Description:

GIS data for the entire State of Indiana, assembled and prepared by experts at Indiana University was added to the data resources provided to the TeraGrid by our partner site Purdue University.

Sharing Information:

Database is available on the TeraGrid to all users as a resource.

Contributions

Contributions within Discipline:

Contributions within the principal discipline(s) of the project;
The principal discipline of this project is the development of large-scale cyberinfrastructure for enabling high-end computational research. To that end, the ETF contributes by enabling researchers to address the most challenging computational problems by utilizing the integrated resources, data collections, instruments and visualization capabilities of nine resource partners.

Contributions to Other Disciplines:

Contributions to other disciplines of science or engineering;
As with any sort of infrastructure project, the contribution of ETF to other disciplines of science and engineering is vast and far reaching and includes the following:
ò Automatic extraction of semantic information from text and links in Web pages and an analysis of their semantic similarity.
ò Literature mining, another area in informatics, to help not only with automatically sifting through huge biomedical literature and annotation databases, but also with linking bio-chemical entities to appropriate functional hypotheses.
ò The analysis of the evolution of hexapods (arthropods with six legs) using a global computing grid.
ò Linked Environments for Atmospheric Discovery (LEAD), bringing advances in cyberinfrastructure tools and techniques to the meteorology community with the goal of enabling more accurate and timely forecasts through on-demand execution of forecast models.
The improved efficacy in the study of X-ray crystallography
ò The real-time analysis of data for weather modeling, bioinformatics, astronomy, fusion energy simulations, and clinical radiation therapy are being developed at IU using MRI facilities.
ò Dynamical simulation studies in astronomy, particle physics, and molecular simulations
ò The development of research methods and tools that allow distributed data and computational resources to be utilized effectively.

Contributions to Human Resource Development:

Contributions to the development of human resources;
Indiana University has strong, ongoing commitments to investing in people and to ensuring that the workforce of tomorrow represents the full richness of American society.

IU coordinates and participates in IT research and education events at the regional and national levels:
ò The annual I-Light Symposium, in conjunction with Purdue University
ò The annual SC conference, the premiere international event for supercomputing, since in 2003, 2004 and 2005.

Through the IP-grid partnership with Purdue and by joining the TeraGrid, IU enhances existing outreach efforts to interest and train people from traditionally underrepresented groups in the study and development of cyberinfrastructure:
ò IU sponsorship of undergraduate intern, Rishi Verma
ò Active participation in several national conferences, including the Grace Hopper Celebration of Women in Computing and the Richard Tapia Celebration of Diversity in Computing.

IU has also committed to significant outreach to many communities in a variety of ways:
ò bringing grid computing information to the HPC community by way of portable, stereoscopic visualization devices
ò sharing grid and high performance computing information with the general scientific community
ò encouraging an appreciation of the global value provided by our HPC and TeraGrid efforts in the lay public of Indiana
ò providing career encouragement in high performance computing to students from kindergarten to graduate school

Contributions to Resources for Research and Education:

Contributions to the physical, institutional, or information resources that form the infrastructure for research and education;
The ETF is, by definition, an infrastructure project. All the resources developed on this project contribute to the infrastructure for research and

education:
ò Networking
ò Computing facilities
ò Visualization tools
ò Massive storage devices
ò Application development

Contributions Beyond Science and Engineering:

Contributions to the other aspects of public welfare beyond science and engineering, such as commercial technology, the economy, cost-efficient environmental protection, or solutions to social problems;
The computational and cyberinfrastructural power of the ETF allows for the development of applications in almost every area of public welfare. Some uses thus far are:
ò tornado prediction
ò drug discovery
ò modeling information processing and political opinion
ò earthquake simulation
ò seismic modeling and oil reservoir simulations
ò groundwater cleanup
ò NanoHUB
ò Identifying brain disorders

Categories for which nothing is reported:

Activities and Findings

1. Describe the major research and education activities of the project.

The TeraGrid is the National Science Foundation's flagship effort to build a national cyberinfrastructure - high performance computers, data sources and massive data storage systems, visualization environments, advanced instruments, and people - all linked by high speed networks. The purpose of the TeraGrid and other cyberinfrastructure is to perform research and create new discoveries and insights that could not be achieved otherwise - that is, without the linkage of many disparate types of systems in different physical locations. The NSF funded Indiana University and Purdue University to join the TeraGrid in September of 2003 through the IP-Grid (Indiana Purdue Grid) project. With the addition of Indiana University and Purdue University through the IP-Grid project, and with the addition of other new resource providers at the same time, the TeraGrid grew to a total of nine resource providers. On October 1, 2004, the TeraGrid was declared to be in production, at the end of a construction phase that had begun, overall, three years before.

This report describes Indiana University's activities building the infrastructure and connections to become part of the TeraGrid as of October 1 2004, and further implementation activities funded through the end of the initial IP-Grid grant. Indiana University and Purdue University have received subsequent funding from the NSF to be Resource Providers on an ongoing basis as part of the TeraGrid via grants 0451237 and 050207.

Indiana University has also taken an active and effective role in the internal organization of the TeraGrid. IU staff (and in particular IU staff funded by base IU funds) have participated actively and reliably in the many working groups, ad hoc technical groups, meetings, and governance activities of the TeraGrid. IU has been particularly active in technical groups involved in portals, storage, accounts management, architecture, and security. In addition IU has made strong contributions within the leadership structure of the TeraGrid - particularly regarding the role of resource providers in the TeraGrid and in shaping plans and objectives for the Grid Integration Group.

While Indiana University’s role in the project is prominent, the TeraGrid would not be successful without all its resource partners including: the California Institute of Technology, the University of Chicago/Argonne National Laboratory, the San Diego Supercomputer Center at UCSD, the Texas Advanced Computing Center at UT-Austin, the National Center for Supercomputing Applications at UIUC, Indiana University, Purdue University, Oak Ridge National Laboratory, and the Pittsburgh Supercomputing Center.

1a) Goals

The primary goal in the creation of the TeraGrid was the construction of a unified and coherent cyberinfrastructure to serve as the foundation of a new generation of computational science activities. To this end, the TeraGrid project deployed a production grid infrastructure that supports high-capability production-level services with grid

middleware technology and policy. This gives scientists the ability to access heterogeneous resources conveniently from multiple sites through a unified set of processes and interfaces including already established science gateways.

The TeraGrid exploits some of the nation's most powerful resources from massively parallel supercomputers to one of the world's fastest networks, from petabyte storage facilities to instruments, visualization engines, and data collections to enable world-class scientific discovery. Collectively, the resource providers supply over 50 teraflops of computing capability, over 1.5 petabytes of online storage, and a selection of specialized resources. The resource providers are physically interconnected by a 40 Gbs national network facility, and integrated operationally through the Grid Infrastructure Group (GIG). The TeraGrid will bring these capabilities to the broad scientific, education, and engineering community through specific partnerships with community infrastructure initiatives, increasing the number of users benefiting from TeraGrid by an order of magnitude within 3-5 years. The infrastructure will improve the productivity of thousands of scientists, providing a persistent framework for all aspects of their investigative research (collaborating with remote colleagues, accessing instruments and data, performing simulations and other analyses, visualizing and archiving their results).

A critical contention in our original proposal to the National Science Foundation was that we form a specific collaboration between Indiana University and Purdue University to jointly become part of the TeraGrid. We have collaborated extensively on all aspects of this project – starting prior to this grant itself when Indiana and Purdue Universities collaborated to create the I-light network (www.i-light.org) which connects the Purdue campus in West Lafayette, the Indiana University campus in Bloomington, and the joint Indiana University Purdue University campus in Indianapolis to each other and to Abilene and the commodity Internet. It was this advanced fiber infrastructure that created the foundation on which both Indiana and Purdue Universities proposed to join the TeraGrid as "IP-Grid" (Indiana Purdue Grid).

The subsequent partnership between Purdue and Indiana Universities as part of the TeraGrid has been extremely productive. Indiana and Purdue leveraged the existing I-light network and created a highly cost-effective, joint 20 GB/sec connection to the TeraGrid backplane, joining at Chicago. Perhaps more important has been the intellectual collaboration between the two universities. The IP-Grid Steering Committee, composed of leaders from the Indiana and Purdue University IT organizations, now meets monthly for sharing of expertise and development activities. For example as Purdue develops the NanoHUB as a TeraGrid portal for nanotechnology and IU develops the Hydra portal as a TeraGrid portal for bioinformatics, IU and Purdue have shared with each other their expertise and lessons learned. These collaborative meetings have also served as a venue to voice common needs within the TeraGrid project to ease the development and addition of such unique resources to the TeraGrid.

While development and testing does take place within the smaller three campus system of Purdue and Indiana Universities, the most serious problems in creating and managing grid technology are problems of scale. Problems that tend to emerge in large scale implementations (e.g. the entire TeraGrid) are not always reproducible in the smaller scale implementations (e.g. three campuses). Therefore part of the initial vision of

collaboration – that the three campus system within Indiana would prove a useful test-bed for TeraGrid technology research and development - has not proven in practice to be particularly useful. However, the collaborations between the computer science researchers and technical implementation teams of the two universities has been particularly fruitful, so the initial concept that Indiana and Purdue Universities could, together, contribute more to the TeraGrid than either could alone, has been born out.

Indiana University and Purdue University have collaborated extensively on outreach and education activities. These range from tutorials presented at several conferences, to the education program at SC05 (the theme of which was grid computing in general and the TeraGrid in particular), to symposia designed to attract new researchers to the TeraGrid. IU and Purdue are responsible for a recurring conference on network-based and grid computing activities called the I-light Symposium. Like SC05, the 2005 I-light symposium was on grid computing in general and the TeraGrid in particular. The 2005 symposium also marked the beginning of a transformation of the conference from a "State of Indiana" event to a more regional conference, drawing attendees from several neighboring states.

This collaboration has proven to be of significant and lasting value to both institutions, the overall TeraGrid project, and most importantly, the national science community.

1b) Establishing network connectivity

Indiana University maintains a connection to the TeraGrid T320 Site Border Router (SBR) in Indianapolis. This router, part of the TeraGrid Autonomous System (AS 75), is managed and operated by the TeraGrid Help Desk. Because of its placement on the Indiana University – Purdue University Indianapolis (IUPUI) campus, the IU Global NOC and engineering staff provide hands and eyes support for the TeraGrid Help Desk. The SBR, providing 10 Gigabit Ethernet connections to Indiana University and Purdue, is connected to the Chicago DTF core node via a pair of 10 Gigabit Ethernet connections. These two connections are provisioned over a Cisco-based Metro DWDM network that rides a pair of fibers acquired from Level3 communications. (See **Figure 1**.)

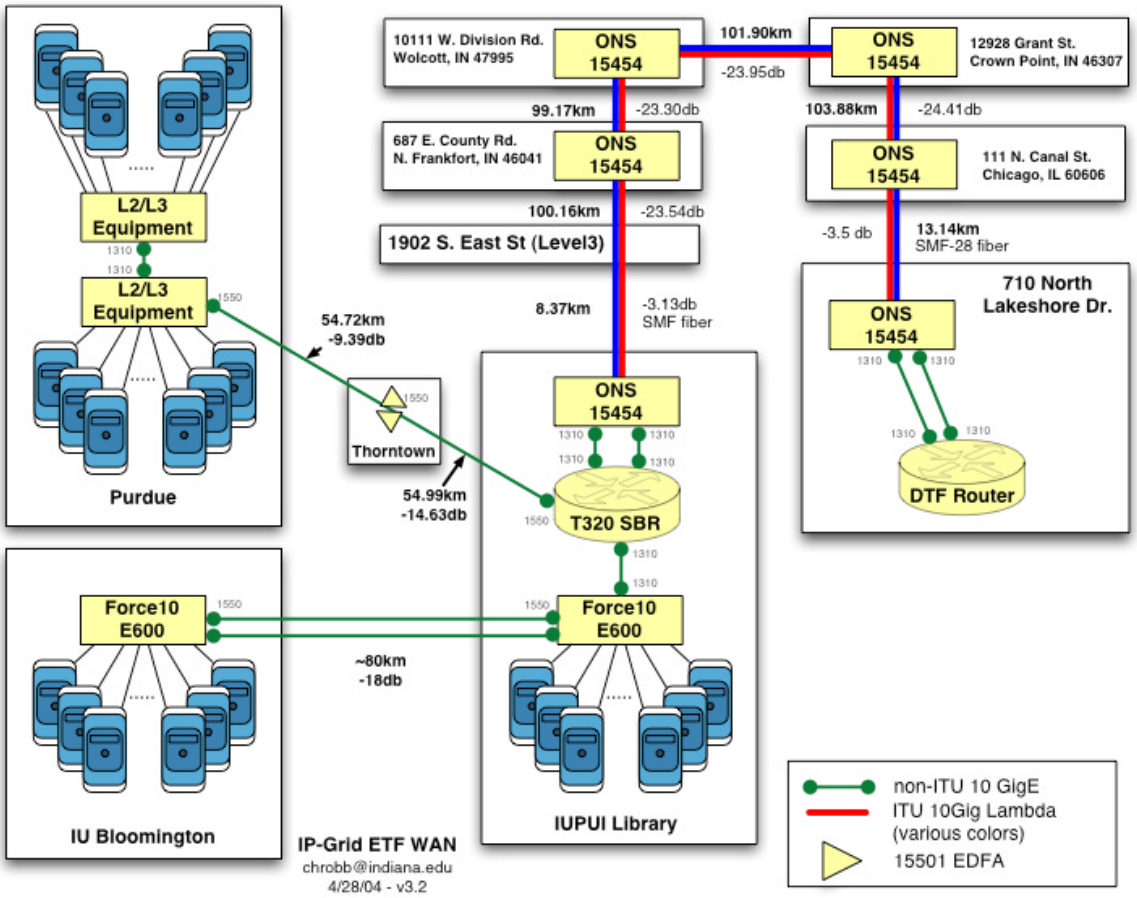


Figure 1: Indiana University connectivity to Chicago

Indiana University TeraGrid Connectivity

Indiana University connects to the TeraGrid SBR via a pair of Force10 E600 gigabit Ethernet switches: one at Indiana University Bloomington and one at IUPUI. Each of these switches provides 120 ports of Gigabit Ethernet connectivity for TeraGrid resources at each campus.

The switches are interconnected via a pair of aggregated 10 Gigabit Ethernet circuits. This inter-campus connectivity is provided via four strands of Non-Zero Dispersion Shifted optical fiber, provided by the I-Light project. Since the distance between campuses is less than 80km, the Force10 switches use extended reach optics to allow for an unamplified signal with no dependence on intermediate optical equipment.

In Indianapolis, the Force10 E600 maintains a 10 Gigabit Ethernet connection to the TeraGrid site border router across IUPUI campus fiber. This connection provides Indiana University TeraGrid resources with a dedicated circuit to the TeraGrid T320 router.

IU TeraGrid Resource Campus Connectivity

Both Force10 switches are connected to the Indiana University campus network via single Gigabit Ethernet connections. This provides the following redundant connectivity:

- TeraGrid resource operational and management capability from the Indiana

University network
- a management path from the Global Network Operations Center at Indiana University to the TeraGrid common equipment
- Abilene access to certain TeraGrid resources as deemed appropriate by the TeraGrid System Management Group
- IU campus access to TeraGrid resources

Campus connectivity is achieved via a separate pair of Juniper M7i campus border routers that are operated independently of the main campus border routers. These routers, currently providing connectivity to just the TeraGrid campus links, provide an extra security buffer between campus resources and campus TeraGrid resources. Network administrators will be able to provide route filtering, packet filtering, rate limiting, or any combination of the three on these campus border routers or on the TeraGrid Force10 switches. The Indiana University campus network engineers administer the border routers. This allows non-TeraGrid campus engineers to effect security or network changes without the possibility of interrupting the access that IU TeraGrid resources have to the TeraGrid. (See **Figure 2**.)

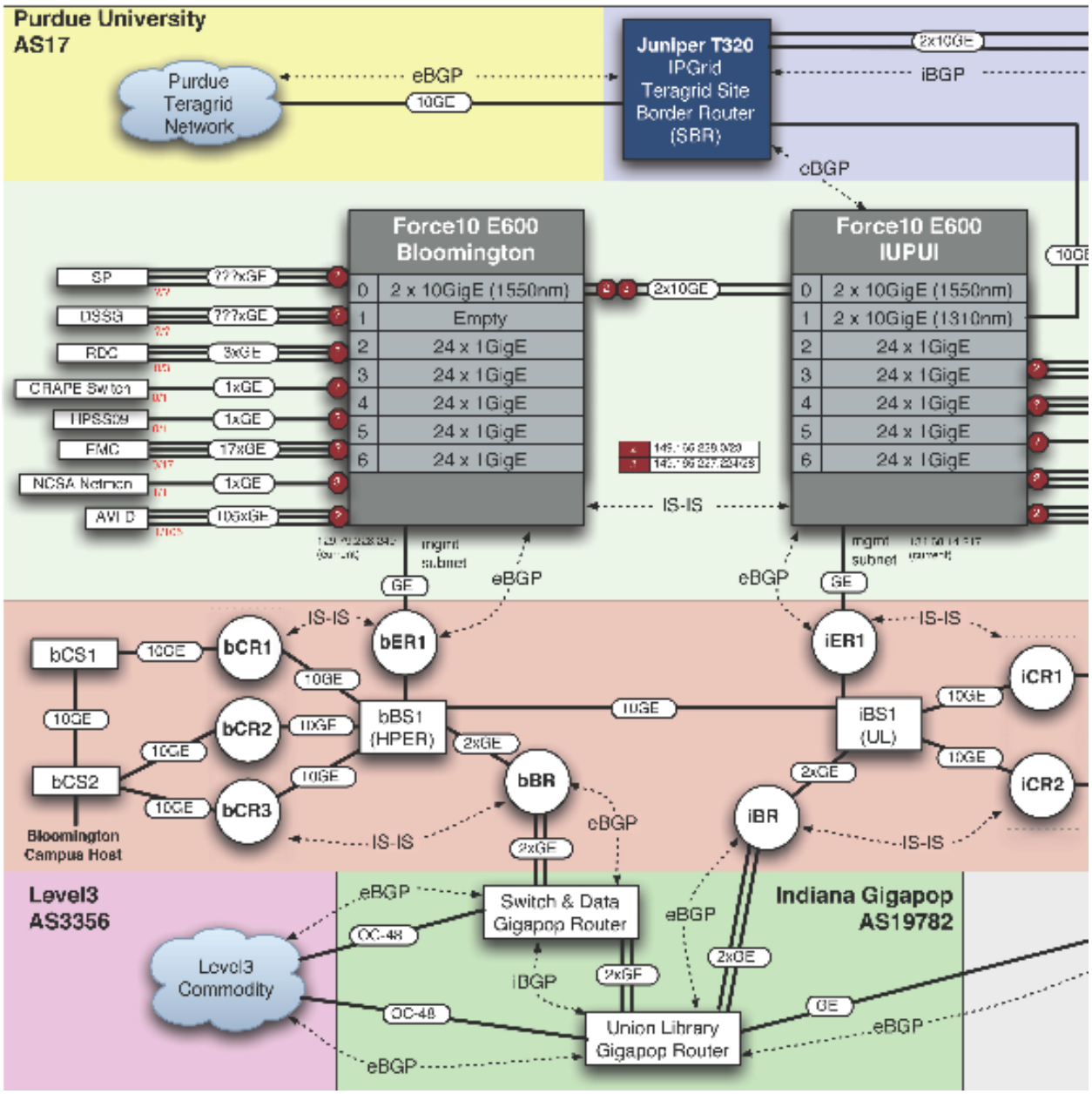


Figure 2: TeraGrid interconnects

1c) Connecting systems to the TeraGrid

Resources

At the end of the grant period, IU had computation resources available to TeraGrid users that were capable of providing greater than 900,000 TeraGrid service units (SUs) annually. Our GRAvity PiPE (GRAPE) boards will be available for the first round of allocations in 2006. Our IBM SP system will be added to the TeraGrid in the 2nd quarter of 2006. Additional Linux resources will also be added later in the year to meet our goal of making 2.3 million SUs available in FY2005/06.

New resources are being provided in other areas as well. Just after the end of the grant period, Indiana University made their HPSS-based Massive Data Storage System (MDSS) available as part of its resource contribution to the TeraGrid, making 1 TB of

storage available on request to researchers with TeraGrid allocations. In 2006, GridFTP access to this storage resource will be made available.

IU's visualization resources are available to TeraGrid users who travel to our Indianapolis or Bloomington campuses. These resources include a high-resolution display wall and configurable virtual reality theater in Indianapolis, the CAVE Automatic Visualization Environment, an Immersive 3D visualization facility in Bloomington, and the portable John-e-Box system, an inexpensive 3D visualization device. (See **Figure 3**.) Plans are underway to make these resources remotely accessible in 2006, including the provision of tools for batch rendering.

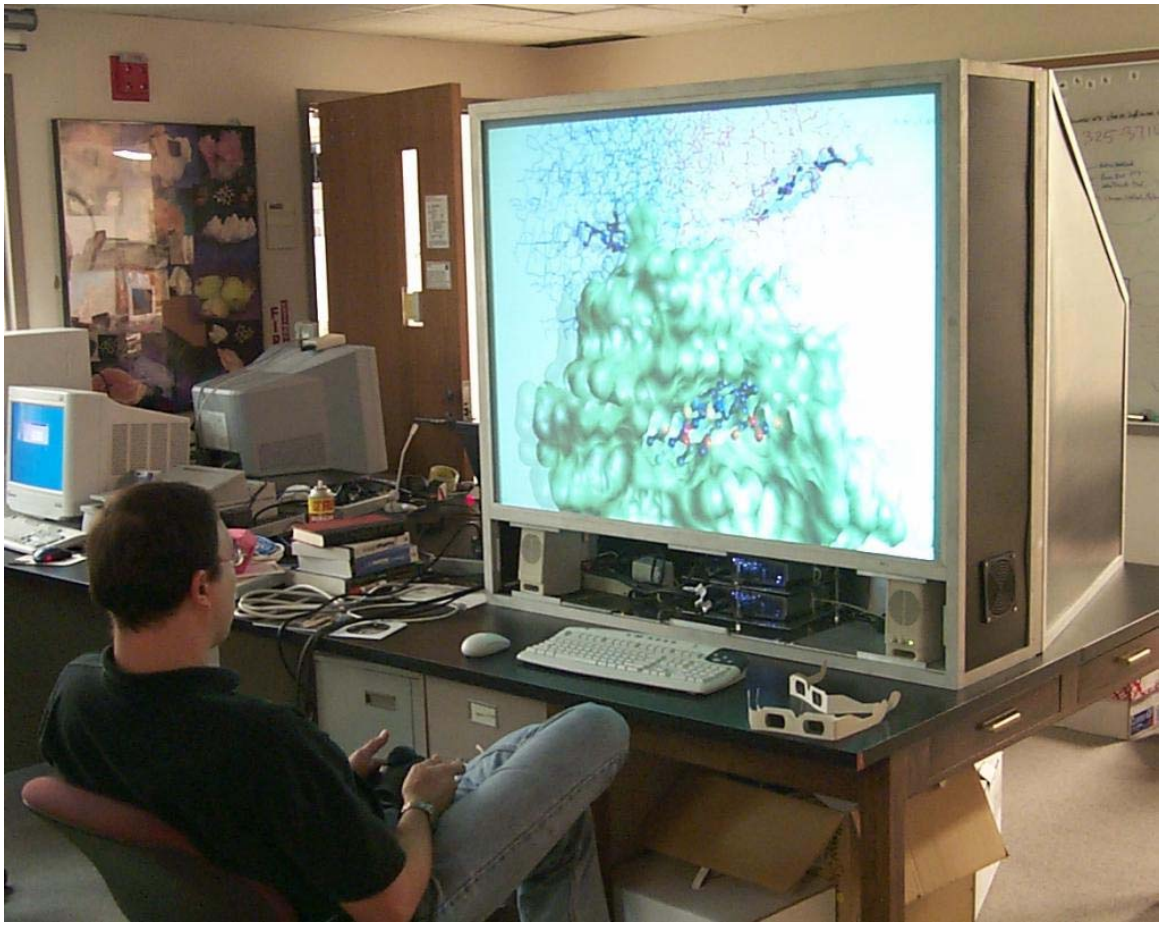
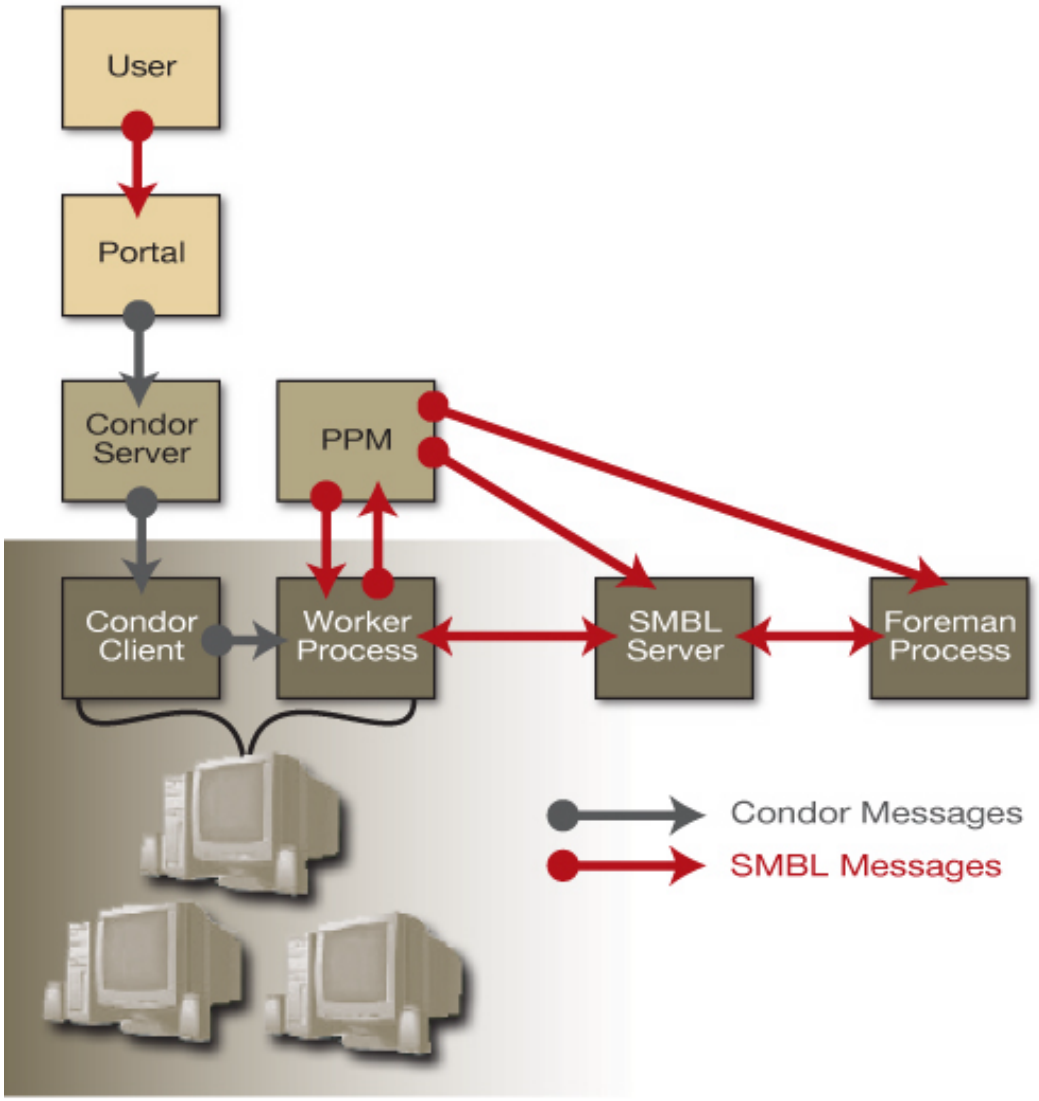


Figure 3: The John-e-Box, a portable, large-format, passive stereo display system, allows researchers to map scientific results in 3D.

In our initial proposal, Indiana University stated that we would leverage Indiana University's award-winning Knowledge Base (kb.indiana.edu) as a support tool for TeraGrid users accessing IU facilities. The use of the KB proved so successful that it has changed the entire support model for the TeraGrid. Indiana University was recently awarded funding under a subcontract from the Grid Infrastructure Group to provide support for the entire TeraGrid via a special instantiation of the IU Knowledge Base. This is one example of the synergy that is constantly found between the universities participating in the TeraGrid, and a particularly important example of IU's participation in that community.

One of the unique resources Indiana University proposed to connect to the TeraGrid is the IU Hydra system. (See **Figure 4**.) This system includes a TeraGrid-compatible Portal as a front end to thousands of computers in student computer labs in use at Indiana University. These computers are harnessed to run three important parallel applications used commonly by biologists: BLAST, MEME, and fastDNAml. These applications are able to run in parallel on a windows-based Condor pool thanks to an IU-developed parallel library called SMBL (Simple Message Brokering Library) and particular programming features implemented by IU in BLAST, MEME, and fastDNAml. SMBL provides a subset of the MPI standard message libraries, and stands between a parallel application and the constantly changing set of resources available to do work in a Condor Pool. This circumvents the usual problem in MPI applications using Condor - the constantly shifting population within MPI_World. SMBL provides what appears to be a constant MPI_World and manages the mapping of these virtual workers to active Condor pool participants. IU has customized parallel implementations of BLAST, MEME, and fastDNAml so that if a Condor pool worker terminates without completing assigned work, the program detects a timeout and resubmits the work to another element in MPI_World. The choice of BLAST, MEME, and fastDNAml for this portal provides thousands of CPUs for use on some of the most important applications in biology.



IU's

Figure 4: Diagram indicating how data passes through a Hydra cluster

AVIDD-IA64 Linux cluster consists of 32 1.3 GHz Itanium2 processors. This resource is 100% dedicated to the TeraGrid and has been available for TeraGrid use since October 1, 2004. It is capable of providing over 275,000 TeraGrid Service Units annually. As part of the AVIDD project, faculty and staff from Indiana University developed the John-e-Box, an inexpensive 3D visualization device. This has enabled installation of 3D visualization devices in individual laboratories and on half of IU's eight campuses.

IU also operates two different types of GRAPE systems - a GRAPE-6, the fastest current system for astronomically-oriented calculations, and two MD-GRAPE systems, designed primarily for Molecular Dynamics calculations. These systems have been heavily utilized in leading edge scientific research.

Linked Environments for Atmospheric Discovery (LEAD) brings advances in cyberinfrastructure tools and techniques to the meteorology community. With a workflow design tool, and a personal user workspace, the LEAD portal allows scientists on the

TeraGrid to easily access atmospheric data, enabling more accurate and timely forecasts through on-demand execution of forecast models. (See **Figure 5**.) It has the added benefit of providing access to high school and undergraduate students to the products and tools of the mesoscale research community.

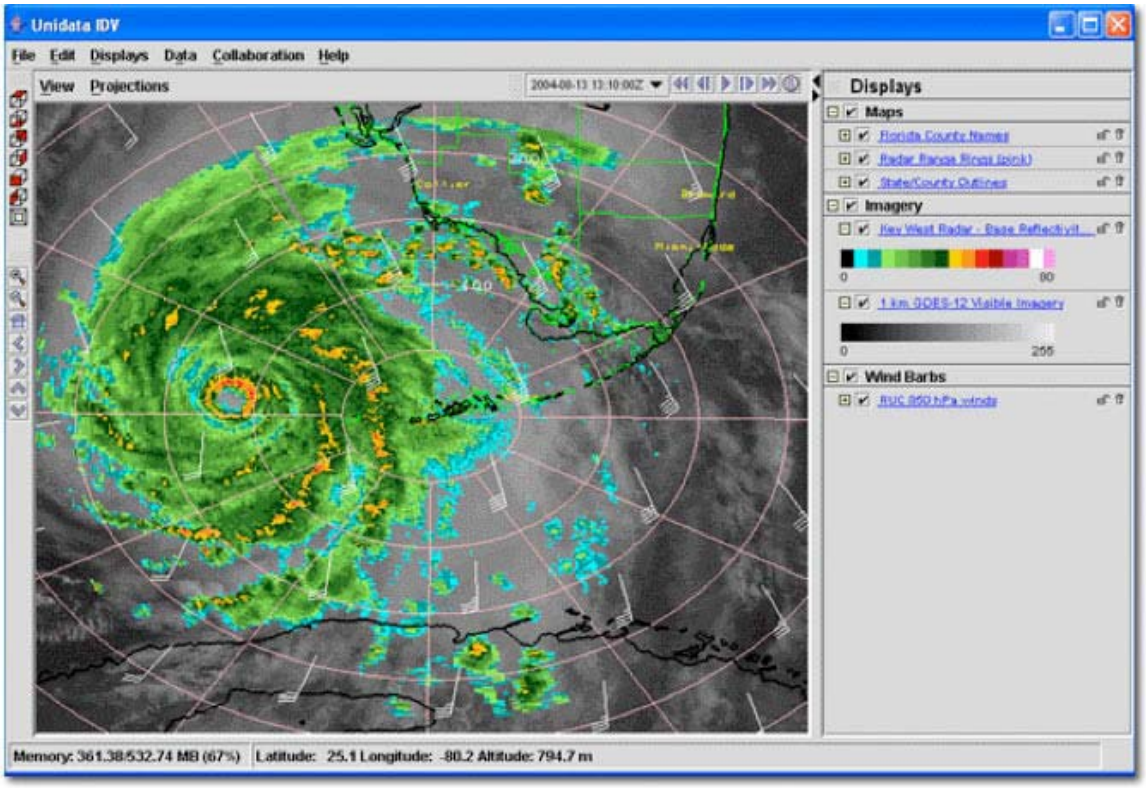


Figure 5: The myLEAD personal information-management tool shows a hurricane off the southwest Florida coast.

The International Virtual Data Grid Laboratory (iVDGL) is where we research and develop research methods and tools that allow distributed data and computational resources to be utilized effectively. The Indiana University Grid Operations Center enhances development and production grid environments by providing information aggregation, common software services and support to users and resource administrators. The Virtual Organization Resource Selector (VORS) acts as an operations tool which allows users to visualize relationships between users with computing needs and computational and storage resources. (See **Figure 6**.)

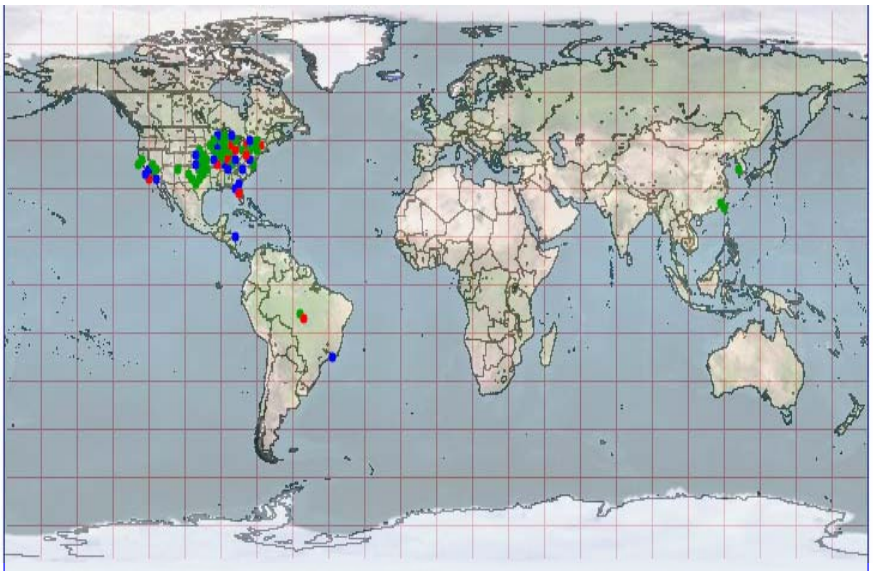


Figure 6: The tool allows users to determine detailed information about the grids, support centers, resources and virtual organizations available for use.

The Data Capacitor is a resource that provides short-term storage – capacitance – to solve three outstanding problems that stem from drawing information from today's large data sets: the need to store data from instruments as rapidly as they can be produced; the need to manage very large data sets for short periods of time; and the need to reliably find and effectively use data available on the Web.

Data Sources

Part of IU's commitment is to provide data sources to TeraGrid users. Some are already available, such as the FlyBase fruit fly genome database. IU's Centralized Life Sciences Data (CLSD) is scheduled to be made available in the 2nd quarter of calendar 2006. The CLSD service provides a single, SQL-based interface for querying a variety of public Life Sciences data, including BLASTable sequence databanks and non-relational datasets that have been transformed into relational tables.

As a resource provider, Indiana's original goals were to focus both on the life sciences and on usability. We are achieving these goals through the delivery of specialized tools such as the Hydra portal and CLSD, through providing user support in the form of the KB, and by supporting the full life cycle of analysis through our focus on storage, data collections and visualization in addition to computing resources.

2. Describe the major findings resulting from these activities.

Indiana University was one of three new institutions that successfully integrated their heterogeneous computing resources into the NSF-funded TeraGrid, and enabled new

computer science and application science discoveries as a result. In conjunction with the creation of new resources in the TeraGrid, many new tools were also developed to facilitate their use.

A key outcome of the Extensible TeraScale Facility and IU's participation in it was the demonstration that this construction and integration project could be completed successfully to the benefit of the national science community.

2a) Tool development

IU's tool development efforts have focused on the following areas:

- Assistance to the TeraGrid in enhancing TeraGrid middleware and accounting tools
- Hydra Portal
- EmbryoGrid
- Linked Environments for Atmospheric Discovery
- Interoperability tools for Open Science Grid and the TeraGrid

The TeraGrid as a construction project has involved more than construction of hardware. It has been a project constructing software as well. Indiana University has participated in the testing, development, and debugging of important software tools for the TeraGrid, including the following:

- Debugging of accounting and monitoring tools
- Implementing the HPSS HSI client so that TeraGrid users can access the IU Massive Data storage System from any TeraGrid Site
- Participation and testing of geographically-distributed GPFS over the TeraGrid.

An accounting system and the tools required for reporting on it are now available for the Hydra portal/condor pool. The Hydra portal is now fully-aware of TeraGrid accounts and will allow users to choose which account to use and then records this to the portal database. Scripts on the back end then generate AMIE-compatible logs that will be provided to the AMIE system. In addition, software libraries, such as the SMBL, and parallel implementations of BLAST, MEME, and fastDNAml have been created for use with the Hydra portal and have been used for projects such as phylogenetic estimation – the process of inferring evolutionary histories by comparing DNA sequence information – enabling scientists to uncover how evolution brought life on earth to its current state.

One of the goals of the TeraGrid is to provide a cyberinfrastructure that supports researchers who need advanced information technology researchers beyond traditional supercomputers, such as data stores. IU is creating EmbryoGrid - a portal for storage of image and video data for the study of embryology - at the request of Dr. Charles Little at the University of Kansas. Dr. Little is a leading expert in using advanced video microscopy to study the migration and division of cells within embryos in development. The data files created by Dr. Little and collaborators at many other universities are very large - we expect several TBs in aggregate - and thus the creation of a embryology data portal to storage resources on the TeraGrid supports important scientific research while broadening the value of the TeraGrid to the US scientific community.

As part of the ongoing support to the Open Science Grid from the iVDGL Grid

Operations area and computational grids in general, the Virtual Organization Resource Selector was produced to allow the user community to easily access resource specific information using a web browser. By filtering the resources displayed depending on the desired Virtual Organization and grid, a user can more quickly find the sites that are available. This tool also breaks new ground by constructing its informational displays based on information from multiple data sources.

Two key tools are at the core of the cyberinfrastructure of Linked Environments for Atmospheric Discovery (LEAD): the workflow design and execution tool, and the user's personal workspace (i.e., "myLEAD"). The workflow tool enables experts and non-experts alike to run complex data ⇒ model ⇒ analysis ⇒ visualization workflows using any of the resources available to the community. The user workspace provides a private space for a user to store model results and a host of other information related to the user's investigations. Access to both private and community data and experiment products are accessible through a visual query interface.

2b) Science Results

Automatic extraction of semantic information from text and links in Web pages and an analysis of their semantic similarity are critical to improving the quality of Web search results. While semantic similarity measures based on taxonomies (trees) are well studied, the design of well-founded similarity measures for objects stored in the nodes of arbitrary ontologies (graphs) is an open problem. Scientists used the AVIDD supercomputer facility to solve a proposed information-theoretic measure of semantic similarity on a graph created from the Open Directory Project (ODP). This graph consisted of more than half million topic nodes, and required more than 5,000 CPU hours on IU's AVIDD supercomputer facility. The computed graph-based semantic similarity measurements, in compressed format, occupied more than 1 TB of IU's Massive Data Storage System. The results from this experiment showed that the new semantic similarity measure improves significantly on the traditional taxonomy-based approach. Results from a data set this large could not be computed without the AVIDD system.

Literature mining, another area in informatics, is expected to help not only with automatically sifting through huge biomedical literature and annotation databases, but also with linking bio-chemical entities to appropriate functional hypotheses. However, there has been very limited success in testing literature mining methods due to the lack of large, objectively validated test sets or "gold standards". To improve this situation, scientists created a large-scale test of literature mining methods and resources, and employed the computational power of the AVIDD clusters in the analysis.

Prior to AVIDD, researchers using X-ray crystallography to identify the structure of chemical compounds often had to discard entire data sets due to unreliable results from a single sample and a lack of recorded environment variables. AVIDD processing capability and visualization tools allow researchers to obtain viable structural analyses that would otherwise not have been possible. AVIDD also enables real-time analysis of data, allowing crystallographers to reduce the time of a run if enough frames have been created for reliable analysis. Similar new capabilities for weather modeling, bioinformatics, astronomy, fusion energy simulations, and clinical radiation therapy are

being developed at IU using MRI facilities.

The MD-GRAPE is a specialized VLSI system useful for any sort of dynamical simulation studies. Indiana University has successfully enabled use of GRAPE systems in astronomy, particle physics, and molecular simulations. IU physicist Chuck Horowitz and colleagues (including one of the UITS staff members involved in the TeraGrid) published a paper regarding the surface structure of neutron stars based on simulations performed using IU's GRAPE-6 system. These calculations would have taken impractically long without the special capabilities of the GRAPE-6 systems.

During the fall of 2003, an international team of computer scientists, biologists, and computer centers created a global grid to analyze the evolution of hexapods (arthropods with six legs). The grid was constituted from systems located in eight countries, spread across six continents (every continent but Antarctica). This work was done as part of the SC03 HPC Challenge, and the project was given an HPC Challenge award for the "Most Distributed Application." The creation of this computing grid enabled investigation of important questions regarding the evolution of arthropods – research that would not have otherwise been undertaken.

2c) Computer Science Research

Linked Environments for Atmospheric Discovery (LEAD) brings advances in cyberinfrastructure tools and techniques to the meteorology community with the goal of enabling more accurate and timely forecasts through on-demand execution of forecast models. The LEAD portal, hosted at Indiana University, allows scientists on the TeraGrid to easily access atmospheric data. Workflow and myLEAD cooperate to provide automatic organization of the user's space, provenance collection, and performance information recording. These tools together promise significant enhancements in the quality of the experience for experts and novice alike in working with the products of computational mesoscale meteorology research.

2d) Usage metrics

Availability and usage statistics for Oct 2004 through Sept 2005 are shown in Figure 7 through Figure 13.

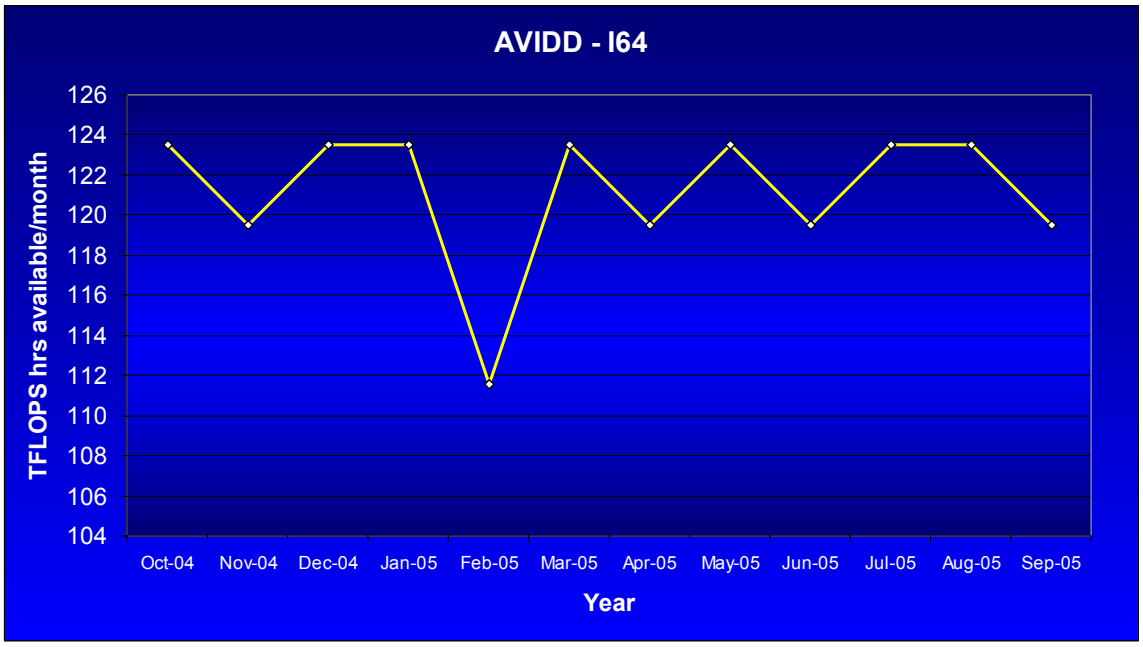


Figure 7: Available TFLOPS hrs for AVIDD-164

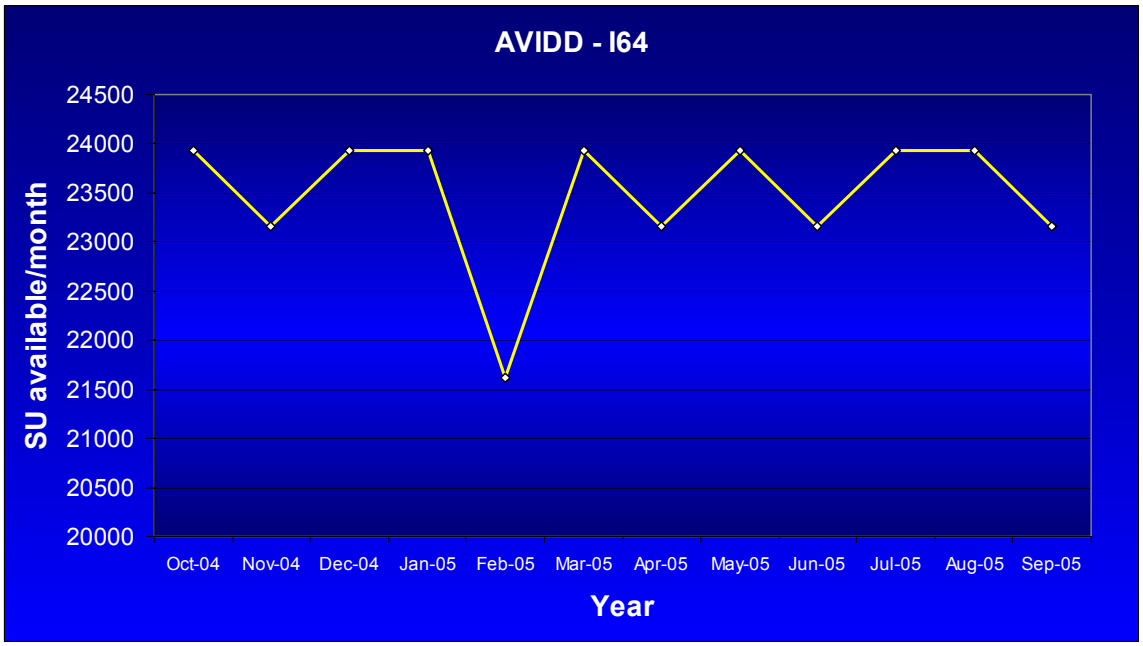


Figure 8: Available Service Units (SUs) for AVIDD-164

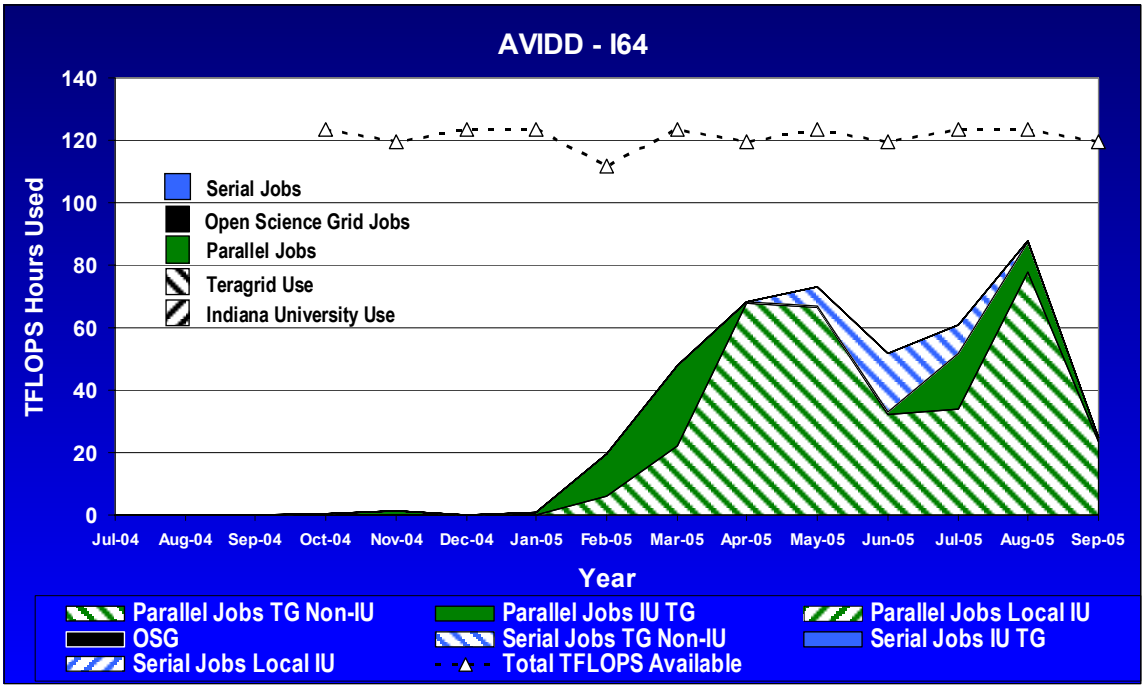


Figure 9: TFLOPS hours used for AVIDD-164

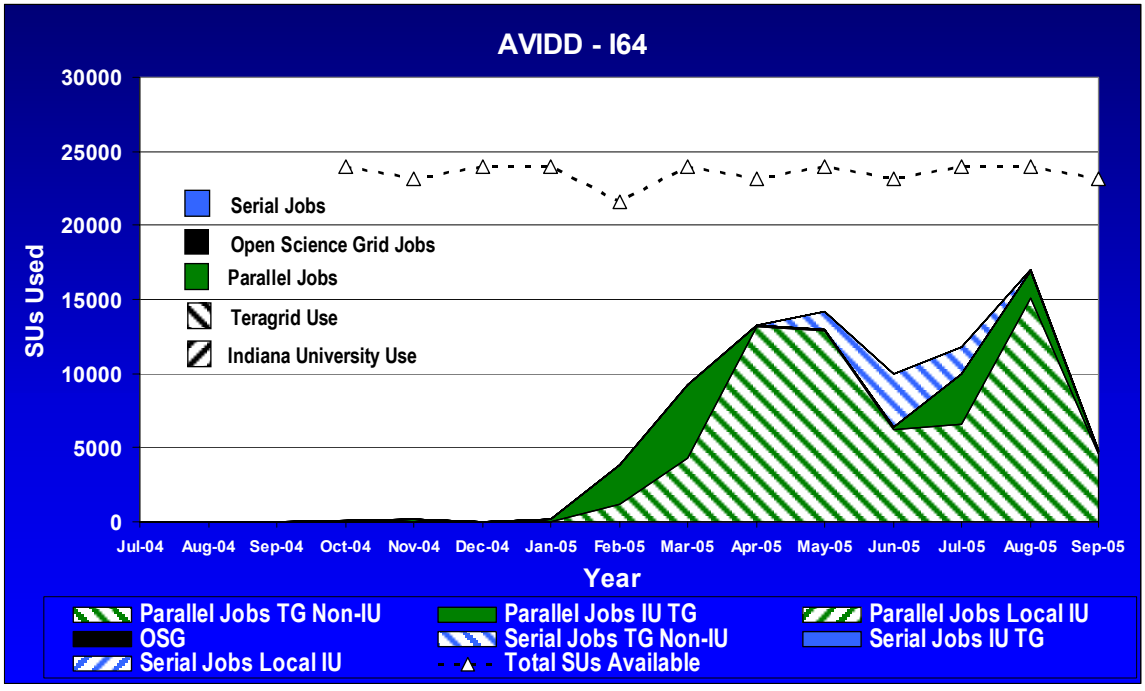


Figure 9: SUs used for AVIDD-164

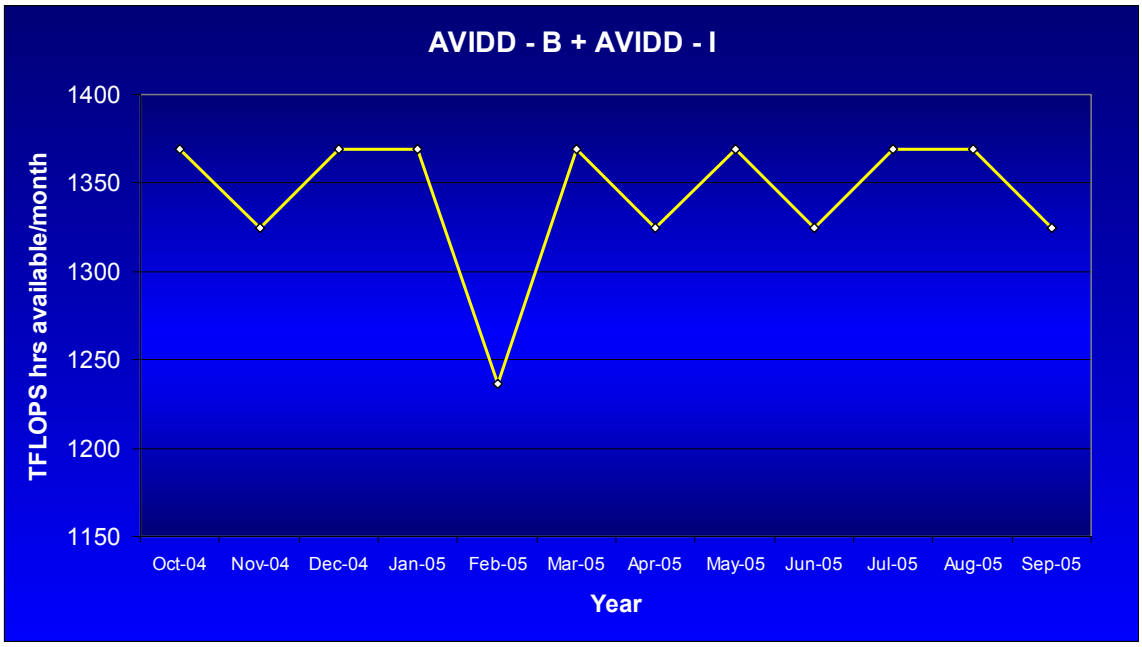


Figure 10: Available TFLOPS hrs for AVIDD-B and AVIDD-1

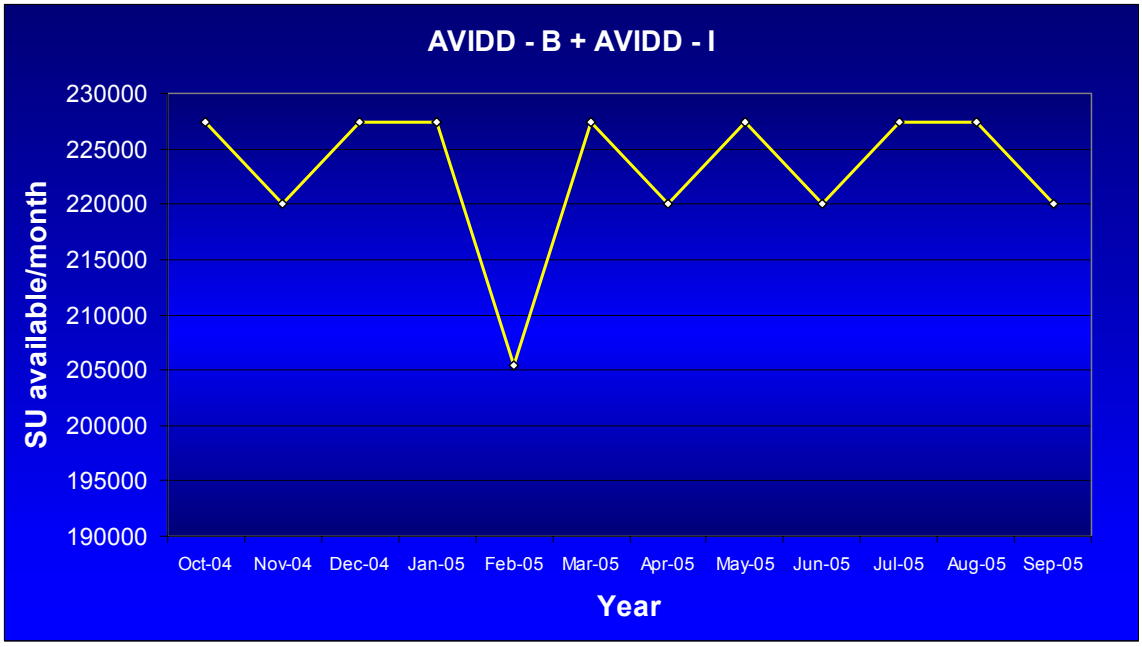


Figure 11: Available Service Units (SUs) for AVIDD-B and AVIDD-1

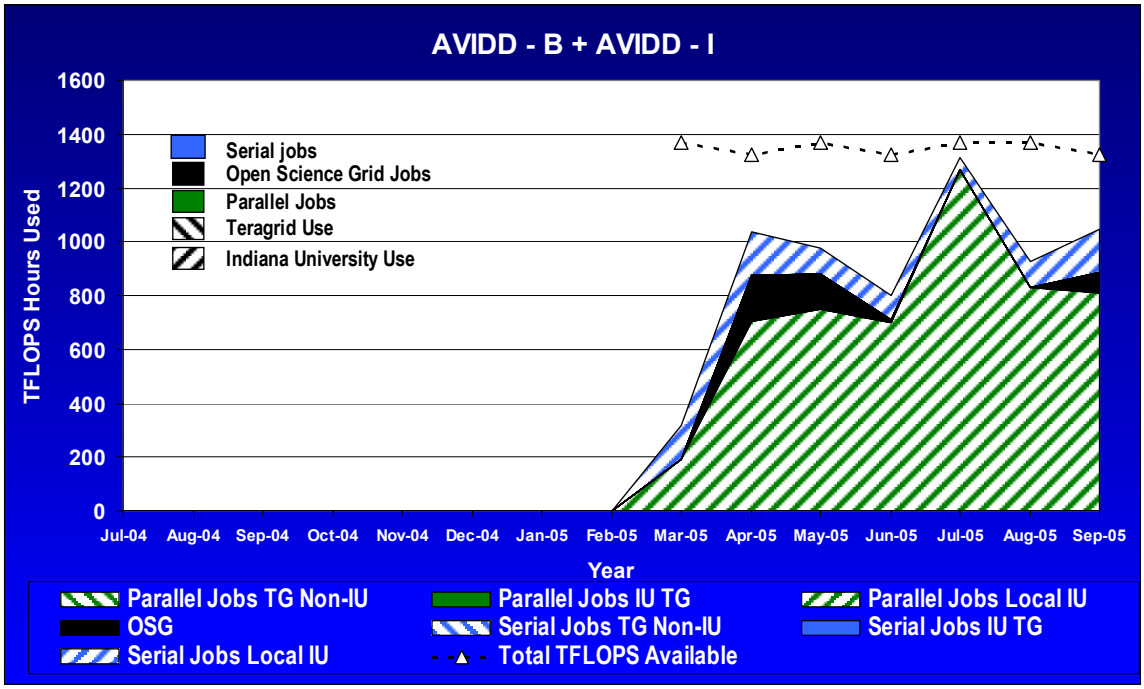


Figure 12: TFLOPS hours used for AVIDD-B and AVIDD-1

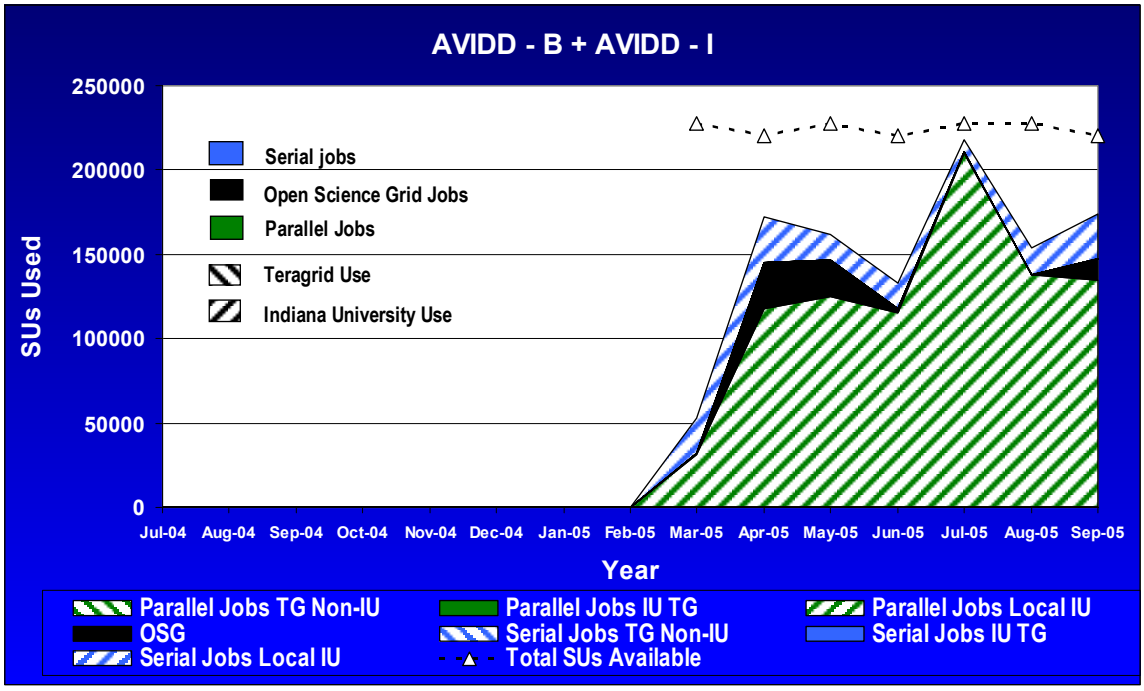


Figure 13: SUs used for AVIDD-B and AVIDD-1

3. Describe the opportunities for training and development provided by your project.

Indiana University coordinates and participates in IT research and education events at

the regional and national levels, disseminating information about the TeraGrid within the advanced computing community. In conjunction with Purdue University, IU has hosted the annual I-Light Symposium (named after the high-speed network connecting the major research campuses in Indiana) to share research and educational applications of advanced networking, computing, and visualization with the broader academic audience across the State. (http://www.iupui.edu/~ilight/symposium05/)

Indiana University has also had a presence at the annual SC conference, the premiere international event for supercomputing, since 1997. In 2003, IU reported on its AVIDD facility, and the formation of the IP-grid. In 2004, IU delivered presentations on the Visualization tools developed for the TeraGrid, the Massive Data Storage System, computer simulations using the MD-GRAPE 2, to name a few. In conjunction with Purdue, IU provided infrastructure support, interactive content, in-depth tutorials, and support staff for the Education Program at Supercomputing 2005. One specific activity provided an opportunity for Randy Heiland, Associate Director of the Scientific Data Analysis (SDA) Lab, one of the Pervasive Technology Labs at IU, to contribute to the SC Education Program. In a show of collaboration, Purdue University researchers, affiliated with the Education Program, invited Heiland to prepare and deliver a Flash-animated, voiced-over Powerpoint presentation. The presentation, "Introduction to Distributed Computing", along with accompanying open source software used by the SDA Lab, was made available on CDs to all teachers who participated in the SC Education Program. This and other online modules were also made freely available through the Sakai project (collab.sakaiproject.org). The SDA Lab actively promotes K-12 science education and outreach (http://sda.iu.edu/K-12)

4. Describe outreach activities your project has undertaken.

IU has strong, ongoing commitments to investing in people and to ensuring that the workforce of tomorrow represents the full richness of American society. Through the IP-grid partnership with Purdue and by joining the TeraGrid, we are enhancing existing outreach efforts to interest and train in cyberinfrastructure people from traditionally underrepresented groups. Results of the research enabled by this proposal will be rapidly disseminated. And our inclusion in the TeraGrid will certainly accelerate nationally funded research underway at our university and magnify the value of NSF efforts in our institution and its two major research campuses.

One IU undergraduate internship was awarded this year to Rishi Verma who worked on a portal to data, including image data and videos, on development of embryos. Mr. Verma has developed a prototype OGCE-compliant portal called "EmbryoGrid" designed to meet the needs of embryologists. Our key contact within the embryology community is Dr. Charles Little, of Kansas State University. Dr. Richard Repasky is the local supervisor of Mr. Verma. We are continuing development of EmbryoGrid as a TeraGrid resource, and hope to have it in general use by 2006.

IU provided support for another undergraduate researcher to learn about supercomputing. Mr. Mathew Burks, an undergraduate, has been doing research in inorganic chemistry since fall of 2004, most intensively during the summer of 2005. IU provided funding for Mr. Burks to accompany the IU team to the SC2005 conference, and has attended several talks on supercomputing applications in chemistry - thus helping interest and prepare him

for use of high performance computing applications in graduate school.

IU has committed to significant outreach to many communities in a variety of ways: bringing grid computing information to the HPC community, sharing grid and high performance computing information with the general scientific community, encouraging an appreciation of the global value provided by our HPC and TeraGrid efforts in the lay public of Indiana, and providing career encouragement in high performance computing to students from kindergarten to graduate school.

For the local and regional communities, IU has utilized the portable stereoscopic visualization devices provided as part of an NSF MRI project (AVIDD, award # 0116050) to present scientific research and educational content to over 2,500 individuals through campus outreach days, IT awareness fairs, and on-site school demonstrations. The UITS Advanced Visualization Laboratory (AVL) was pleased to participate in the Brownsburg Challenger Center's 10th Anniversary celebration held in November 2004. During this all-day event, AVL staff showcased 3D space-related content to nearly 100 visitors including children of all ages and their parents. (See **Figure 14**.) Of particular note was the summer 2004 partnership with the Indianapolis-Marion County Public Libraries that toured eight branch libraries in different demographic regions as part of the "NASA in Your Library" program. More than 450 people, mostly school children, attended these demonstrations. (see **Figure 15**.) One portion of the distributed AVIDD cluster, and one of AVIDD's 3D visualization systems, was installed on the IU Northwest campus located in Gary, Indiana. As a result, many students from traditionally underserved groups were able to use advanced scientific computing and visualization systems.

Figure 14: Held on the IUPUI campus annually, "Explore IUPUI" is an opportunity for the community, young and old, newcomers and old friends to see what goes on at IUPUI. The AVL showcased select projects including student animations, NASA images, and a fly-through of the Indianapolis downtown.

Figure 15: The Advanced Visualization Laboratory (AVL) of UITS/RAC put on demonstrations of stereoscopic (3D) visualizations at seven different branches of the Marion County Public Library System in conjunction with their summer reading program.

IU has also sponsored and actively participated in several national conferences focused on traditionally underrepresented groups, including the Grace Hopper Celebration of Women in Computing and the Richard Tapia Celebration of Diversity in Computing. (IU was a Bronze sponsor for these conferences in 2004 and 2005.) Thanks in part to this sponsorship and a very active Women in Computing organization (WIC@IU), IU had 14 representatives at Grace Hopper 2004. IU researchers also presented results at Tapia 2005. The WIC@IU group has also developed an interactive experience called "Just Be" that seeks to break common stereotypes about people in computing. With support from UITS, this program has been presented at a number of middle schools and high schools in Indiana, Kentucky, and Missouri.

We have actively promoted the TeraGrid to technical and educated lay audiences through presentations and talks. A list of talks is provided below:

Shankar, A.. 2004. "The Future of Computing" presented at the Tata Institute of Fundamental Research, Mumbai, India.

Shankar, A.. 2004. "The Future of Computing" presented at the Grid Workshop, Center for Development of Advanced Computing, Bangalore, India.

Shankar, A.. 2004. "The Future of Computing: IU and the TeraGrid" UITS IT Seminar Series, Indiana University, Bloomington, IN.

Shankar, A.. 2004. "The Future of Computing" presented at the National Center for Software Technology, Bangalore, India.

Shankar, A. and Verma, R.. 2004. "TeraGrid: Laying the Seeds for Tomorrow's Computing" presented at Bloomington High School South, Bloomington, IN.

Shankar, A.. 2004. "The Future of Computing: IU and the TeraGrid" UITS IT Seminar Series, Indiana University Purdue University at Indianapolis, IN.

Shankar, A.. 2004. "The Future of Computing" presented at the Indian Space Research Organization, Bangalore, India.

Shankar, A.. 2004. "The Promise of Grid Computing" presented at the Indian Institute of Technology, Delhi, India.

Shankar, A.. 2004. "The Promise of Grid Computing" presented at the Center for Development of Advanced Computing, Pune, India.

Shankar, A.. 2004. "The Future of Computing" presented at Jawaharlal Nehru University, Delhi, India.

Shankar, A.. 2004. "The Future of Computing" presented at the Center for Development of Advanced Computing, Mumbai, India.

Shankar, A.. 2004. "The Promise of Grid Computing" presented at the Ministry of Communications and Information Technology, Delhi, India.