A History of the TeraGrid Science Gateway Program: A Personal View

Nancy Wilkins-Diehr
San Diego Supercomputer Center
9500 Gilman Dr.
La Jolla, CA 92093-0505
001 858-534-5118
wilkinsn@sdsc.edu

ABSTRACT

This paper describes the NSF TeraGrid Science Gateways program, its formation, progress, lessons learned and current contributions over its seven-year life and new directions in the NSF XSEDE program. Early requirements analysis work with path-finding gateways that formed the underpinning of the program are described as are current projects and their unique contributions to the larger program. Future directions both within the XSEDE program and for gateways more generally are discussed.

Categories and Subject Descriptors

A.1 Introduction and Survey

General Terms

Management

Keywords

Science gateways, portals, computational science, middleware

1. BACKGROUND

The Mosaic browser was first developed at NCSA in 1992 [1]. By the beginning of the next decade, web technologies were beginning to advance rapidly from static HTML to interactive, dynamically generated web pages that could communicate with other systems. When Mosaic was released in 1993, there were 10 million worldwide users of the internet. Today there are 1.8 billion [2]. At the same time, scientific data was becoming increasingly digital. Computational simulations, collections of digital data from sensors and instruments and publicly contributed digital data were all growing substantially. It was at the beginning of this trend that the TeraGrid Science Gateways program was developed. It is the author's hope that a reflection such as this can be helpful for others developing similar programs.

Accomplishments over the years have been significant. In 2006, 10 gateways used fewer than 100,000 CPU hours across the TeraGrid In 2010, 22 gateways used 40 million hours. By 2011, nearly 40% of users charging jobs on the TeraGrid came through gateways, led by the CIPRES phylogenetics gateway [3] which accounted for 25% of all TeraGrid users.

Lessons learned are identified throughout the paper and are summarized in Section 7. Throughout the paper, team member contributors are identified in parentheses.

2. EARLY YEARS

Supercomputing had always been a rather individual undertaking. Small numbers of individuals from research groups with high-end computing experience got accounts, ran programs at the command line through batch queues, analyzed the results, sometimes with simple visualization programs and submitted more programs to run.

In 1999 NSF launched its first Information Technology Research (ITR) program [4]. A President's Information Technology Advisory Committee (PITAC) report [5] released earlier had cautioned that funding for information technology research was "dangerously inadequate". This first ITR program called for research spanning information technology and scientific applications. As these early projects developed, it became clear that the application of high-end computing resources could have a significant effect on analysis capabilities.

The initial Science Gateway program was envisioned by Rick Stevens at Argonne National Laboratory in 2003 while developing the NSF Extensible Terascale Facility (ETF) proposal for the TeraGrid. The first ITR projects were reaching maturity and seemed a natural fit for the fledgling gateway program. In the proposal, science gateways were part of the TeraGrid's "wide" strategy to have an impact on more individuals and more research groups. From that original proposal:

> "Despite the technological progress of grid technology and deployment, only a minority of the scientific, engineering, and education community use today's national computing infrastructure. Our WIDE strategy addresses this situation by working directly with specific community leaders who are building discipline-specific cyberinfrastructure capabilities and resources for their communities. Our "science gateway" initiative has two avenues for coupling TeraGrid to community infrastructure. The first is a set of application "service" interfaces, identified by partner communities, which will support the use of TeraGrid services through existing applications or custom web portals. The second is a set of technical and policy efforts with peer grids, beginning with the Open Science Grid (OSG), to integrate TeraGrid resources into those peer grids based on the needs of their associated user communities. This integration will permit resource sharing and scheduling of jobs on TeraGrid via existing peer grid infrastructures."

Ten "partner communities" were brought in to help define TeraGrid's efforts to develop common protocols, technology and middleware for gateways, specifically simplified access to job submission, data movement, workflow tools and data-streaming protocols.

These projects and their science leads were:

- Linked Environments for Atmospheric Discovery (LEAD) (Droegemeier, U Oklahoma) [6]
- National Virtual Observatory (NVO) (Szalay, JHU) [7]
- Network for Computational Nanotechnology (Lundstrom, Purdue) [8]

- National Microbial Pathogen Data Resource (Stevens, U Chicago/ANL) [9]
- Building Biomedical Communities (Reed, UNC) [10]
- Neutron Science Instrument Gateway (Cobb, ORNL) [11]
- Grid Analysis Environment (Newman, Caltech) [12]
- Emergency Decision Support (Eubanks, LANL)
- Real-Time Urban Flood Hazard Analysis System (Urban, U Texas)
- Open Science Grid (Pordes, FNAL) [13]

The proposal included one-page project descriptions with a 5-year timeline of deliverables and estimated increases in user counts for each of the projects. These user counts turned out to be vast over-estimates, approaching 5000 users at the end of 5 years.

The proposal did state that "As these gateways are implemented and enter production service, they will serve as examples for which we will develop guidelines and primers to facilitate additional science gateway communities. By years 3–4 we expect to move some of the gateway prototyping staffing into a support infrastructure for the growing number of science gateways."

Our prototyping work began with a survey of the ten initial projects.

2.1 Requirements Analysis Team

One hallmark of the TeraGrid program is the formation of requirements analysis teams or RATs to pursue activities that span more focused "working groups" such as accounting or security. The Science Gateway RAT, led by Sebastian Goasguen was the first of these many RATS. RAT members included John Cobb, Dennis Gannon, Jeff Gardner, Eric Roberts, Von Welch, Nancy Wilkins-Diehr and Roy Williams.

Extensive interviews with each of the 10 projects were conducted over 2 months. A report summarized recommendations for 7 TeraGrid working groups - accounting, security, portals, software, user services, scheduling and data. Summaries of individual project interviews were included in an appendix.

This initial work would serve as the foundation for the entire gateway program. The goal was to develop a common set of services offered by TeraGrid resource providers so the gateway program could be operated in a scalable manner without developing unique services for each project. Early in the program it wasn't clear that this could be done.

Gateway developers identified the need for different types of accounts - developer, student, high-end researcher - with varying capabilities. This need would serve as the foundation for community accounts, the restrictions imposed on these accounts, and the need for gateways to do their own accounting and auditing to keep track of individual usage. Dynamic, short-lived accounts used by the TeraGrid Visualization Gateway were also explored as a way to provide this varied access, but this approach never got wider traction. Some gateway teams explored strong and weak authentication similar to the vetted and unvetted user approach being put in place in the XSEDE User Portal today. Some questioned whether TeraGrid ought to run certificate authorities for gateways, but we saw little demand for this. Policies regarding certificate acceptance found a home in the security-wg and were found useful by command line users as well as gateway developers.

Early on developers expressed interest in a metascheduler, a single job submission interface to any TeraGrid resource. Though there were interfaces that reached a subset of platforms and tools to estimate queue waits, there has never been a single queue interface to all resources. Some in the survey described chained workflows, on-demand scheduling and co-scheduling (scheduling two machines simultaneously) as needs. Co-scheduling lessened in importance as machines grew more heterogeneous and as single machines grew larger. On-demand scheduling is another example of a capability where, if not obviously available, will not attract the relevant user communities.

Web services were described in TeraGrid annual reports as being in their early definition phase when the gateway program began. Standards were being defined by entities such as Microsoft, IBM, Sun, HP, Oracle and the standard bodies like W3C, Oasis and the then-named Global Grid Forum (now Open Grid Forum). 5 of 9 initial gateways interviewed expressed a need for web service interfaces. There was a push to move to Globus 4, with its WSRF interfaces, but this type of usage did not pan out widely with the gateways. Perhaps because of the lack of demand, TeraGrid never did develop widely used web interfaces to standard tasks like job submission and account management.

Since our experience with portals was so new, we also surveyed projects for the type of software they were using for portal development. Technologies included GridPort [14], Clarens [15], In-VIGO [16], OGCE [17] and GridSphere [18]. Some discussed leveraging TeraGrid User Portal technologies, though this never came to pass directly. User portal staff were, however, strong contributors on gateway teleconferences. Some requests foreshadowed the need for the Quarry virtual machine capabilities offered years later. Some requested that specialized software become part of the Common TeraGrid Software Stack (CTSS). This was resisted in that there was very much the desire not to have individual software requirements for each gateway, but rather have a standard that would meet the needs of many.

We identified the need for policies for "anonymous" portal users. This grew into the process for requesting community accounts, which now includes requirements for contact information and estimates of disk and CPU usage requested by TeraGrid's security working group so that unexpected behavior can be detected. Gateway developers were advised to keep auditing and accounting records so that individual use could be identified. In early discussions, the security working group requested mechanisms to identify and restrict problem jobs submitted by individual gateway users. Tools to do this, however, were never fully developed and gateway developers were comfortable with all jobs being shut down if problems were identified with a community account. In practice, this streamlined approach worked well because of the low incidence of community account related security problems.

Initial surveys indicated a need for a standard set of ports to be open for gateways and that running a certificate authority server for a gateway would be useful offerings, but this turned out to not really be the case.

We discussed the need for certificate acceptance and while such policies were developed, this did not turn out to be a gateway-specific issue. We also discussed export-controlled codes because of early work with a Homeland Security gateway that never materialized. While the issue of export-controlled codes has not re-surfaced as an issue yet, international usage of gateways remains an issue.

The portal working group was asked to study Clarens [15] and In-VIGO [16] and also to be aware of auditing and accounting requirements from the security working group. The software working group was asked to compare the TeraGrid and OSG infrastructure for overlaps. A special OSG RAT was recommended to look at cross-grid job submissions.

The community software area concept, first defined in the Gateway RAT, allowed users to move beyond their home directories to install codes that benefit a community in a production filesystem with backups and larger disk quotas. The survey also identified the need for initial gateway documentation, instructions on uploading data sets and developing web services interfaces to these datasets. However since most resource providers did not support web interfaces to their filesystems this need had to be addressed in other ways. Some gateways accessed data on TeraGrid through interfaces like gridFTP and then made the data available through a locally-hosted gateway.

A gateway primer [19] was recommended so that other groups could benefit from the infrastructure being developed. An outline of such a primer, covering many of the areas listed in this section was completed in 2006 and served as the basis for future documentation and tutorials.

3. TRANSITION TO PRODUCTION

The initial plan for the gateway program underestimated the pace of progress. TeraGrid staff were able to move beyond prototypes and put production infrastructure, processes and policy in place well before years 3-4. The 5-year road maps from the proposal, however, reduced flexibility somewhat and slowed the pace of this transition as activities were reassessed.

Some of the initial ten projects didn't develop because of lack of source code access (Urban Flood Hazard Analysis). Others didn't develop as originally envisioned (Emergency Decision Support), but led to other capabilities such as the development of SPRUCE (Special Priority and Urgent Computing Environment) [20], which provides an authentication layer for on-demand computing. Still others (NVO [7]) challenged us to provide services unavailable at the time (gateway hosting) that are now a well-used part of the infrastructure.

Two years into the project, we were able to transition from the staffing of prototype projects to providing support for the larger gateway user community, including those requesting such support through the TeraGrid peer-review allocation process. We felt infrastructure and services had reached a production state and were ready for a larger gateway user community. A common set of requirements for gateways had been identified, often with overlapping benefits for command-line users. This transition in the approach to support was actually a turning point in the program. Funding for gateways, as well as the responsibility for primary development and user needs assessment once again came directly from the science community. The TeraGrid program was not developing gateways for others, but rather helping others with the effective integration of high-end cyberinfrastructure. Because staff had been involved in many successful projects and because the team had experience with a large number of projects coming through the program, we were often in a position to make recommendations regarding development, however development rested with those closest to the science community. This approach has contributed greatly to the scalability of the program. However sustainable funding for gateways from the science communities can be problematic.

Outreach and talks about the program resulted in a steady stream of interested communities. Some examples include the Geosciences Network (GEON) summer institute, the National Biomedical Computation Resource (NBCR) summer institute, Humanities, Arts, Science, and Technology Advanced Collaboratory (HASTAC), CI Channel, the Lariat Networking Project, Minority Serving Institutions CI Empowerment Coalition (MSI-CIEC), Southeast Universities Research Association (SURA), Grace Hopper conference and a European Geophysics Union (EGU) meeting.

4. PROGRAM HIGHLIGHTS

Perhaps a hallmark of the TeraGrid gateway program has been the leadership staff members showed and continue to show in the community. While the term science gateways itself originated in the TeraGrid program, it is now commonly used throughout the world. There are international workshops on science gateways, books on Amazon about gateways and special journal issues on gateways (Figure 1).

Figure 1 Select activities highlighting science gateways.

The Gateway Computing Environments workshop series (first named Grid Computing Environments) has been a featured workshop at the SC conference since 2005. This series, organized by Marlon Pierce, features the work of gateway developers throughout the world. TeraGrid gateway staff have often served on the program committee as well as contributing technical papers.

The gateway program was introduced internationally at a DEISA (Distributed European Infrastructure for Supercomputing Applications) meeting in Paris in April, 2005, followed by an international workshop held at Global Grid Forum 14 (June, 2005) that resulted in a special journal issue of *Concurrency and Computation: Practice and Experience* [21]. A new series has since emerged from Europe entitled the International Workshop on Science Gateways (IWSG) with workshops held in Sicily (2010), London (2011) and planned for Amsterdam (2012).

Gateways have been popular in the classroom as well. Gateways were featured at the NVO summer school (Williams, 2006), at a LEAD workshop for Unidata atmospheric scientists (Christie, Marru, Pierce, 2006), and as part of a 10-week Weather Challenge (Christie, Marru, Pierce, 2007) involving students from 10 universities running weather simulations to see who could most accurately predict the weather. The Weather Challenge imposed a significant load on the TeraGrid and uncovered reliability problems. While the Weather Challenge was completed successfully, the effort to achieve this led to an in depth debugging exercise involving grid software developers, gateway developers and TeraGrid system administrators. This resulted in reproducible tests to check stability, changed the way grid software was installed and operated and eventually changed the course of the software development itself. Gateways were also used in problem sets for a student competition at TeraGrid 07 and in CASP9 protein structure prediction challenge sponsored by the US National Library of Medicine. The Open Protein Simulator (OOPS) gateway [22] was used for the CASP9 (Uram, 2010).

Activities continued to further community interest in gateways. Presentation were given to all NSF directorates (Wilkins-Diehr, Klimeck, Gannon, Pamidighantam, McGee, Wang, May, 2007), at the Canadian Cyber Summit (Wilkins-Diehr, October, 2007), at the US-China Cyberinfrastructure Symposium in Beijing (Wilkins-Diehr, April, 2009), at the EGEE conference in Barcelona (Wilkins-Diehr, September, 2009) and at a keynote presentation at IWSG 2010 in Italy (Woitaszek).

A CyberGIS workshop at the Congressional Reading Room, Library of Congress (Wang, Wilkins-Diehr, February, 2010) highlighted the potential for gateways in Geographic Information Science. This gave the gateway program some high level exposure. Administration attendees included Karen Siderelis, GIO Department of Interior, Acting Chair Federal Geographic Data Committee; Raphael Bostic, Assistant Secretary for Policy Development and Research, HUD; Derek Douglas, Special Assistant to the President for Urban Affairs; Jerry Johnston, CIO, EPA and Stephen Lowe, GIO, Department of Agriculture.

Publication highlights related to gateways include an IEEE Computer article "TeraGrid Science Gateways and Their Impact on Science" [23], *Virtual Research Environments: From Portals to Science Gateways* [24] available on Amazon, a chapter in CRC Press' *Cloud Computing and Software Services: Theory and Techniques* entitled "Science Gateways: Harnessing Clouds and Software Services for Science" [25] and a cover image in the Proceedings of the National Academy of Sciences [26].

More operational activities included a build-a-gateway in a day tutorial (Shankar, Sharma, Martin, Mock, Liu, Wang, Wilkins-Diehr, June, 2007) that featured SimpleGrid [27]. SimpleGrid condensed a complex gateway into its simplest moving parts and served as a good blueprint for many others beginning gateway development. We found this important to offer as many of the initial projects we worked with were quite mature and it was difficult for new developers to understand the underlying components.

In January 2008 the Gateway Security Summit (Wilkins-Diehr, Marsteller) brought together gateway developers and security experts to discuss potential risks in a supercomputing environment due to gateways and options for providing additional security. Finally in 2011, a formal TeraGrid policy was approved recommending standardized treatment of community accounts across all resource providers (Hazelwood, Woitaszek, Wilkins-Diehr, May, 2011).

5. CURRENT PROJECT WORK

In this section I provide an overview of current projects where TeraGrid and now XSEDE staff are providing labor support. I highlight the unique contributions each project brings to the gateway program. This type of diversity strengthens the program and allows developers in often disparate fields to learn from each other. Names in parentheses are staff members responsible for a given project.

CIPRES [3]. (Miller, Pfeiffer, Schwartz) The Cyberinfrastructure for Phylogenetic Research (CIPRES) gateway is a "public resource for inference of large phylogenetic trees" and has the largest user community using TeraGrid of any gateway. In recent quarters, 25% of all users charging jobs on the TeraGrid did so by using CIPRES. Users come from all continents except Antarctica and from 17 EPSCoR states in the US. CIPRES is used in many classes, but also used for major research resulting in publications journals such as Nature, Cell and the Proceedings of the National Academy of Sciences. The success of CIPRES has helped shape allocation policy and supporting infrastructure for other gateways, for example the ability to charge individual allocations through a gateway.

GridAMP [28]. (Metcalf, Woitaszek) The Grid Asteroseismology Modeling Portal (AMP) gateway team provided a nice model for early requirements gathering that led to a successful gateway deployment that met the user community's needs. The team started with international workshops to poll scientists and a NASA science review. The work is tied to data streams from the Kepler satellite. Users run very compute-intensive jobs, primarily on Kraken and came from 8 countries in recent quarters. This team has served as beta deployers of several important TeraGrid capabilities - attribute-based authentication, commsh and GRAM5. They have also pioneered job chaining on Kraken, greatly reducing queue waits. Their streamlined approach to gateway development enabled an entirely new group to build a gateway after a 20-minute telephone conversation and a review of documentation.

PET. (Woitaszek) The Population-Environment-Technology model (PET) gateway work is just beginning, but promises to expand usage to very new communities with potentially dramatic results. PET is a global economy model that simulates the action of various sectors of the economy for regions of the world. Economic inputs and outputs are balanced in a region and between regions through trade, given pre-specified growth rates of things like population. This type of gateway makes it possible for people to experiment with how changing parameters changes the behavior of the economy. For example researchers can investigate how changing the interest rate, the savings rate, the capital tax rate, or introducing carbon taxes in a certain region affects sectors of the economy such as energy resource use and capital savings.

SIDGrid [29]. (Kenny, Uram) The Social Informatics Data Grid (SIDGrid) gateway has been pioneering in many areas. It is one of a small number of social science applications on the TeraGrid and the only social science gateway today. It provides access to expensive, unique, multi-modal datasets which include streaming data that change over time - voice, video, fMRI images, text annotations and numerical data (heart rate, eye movement, etc.). SIDgrid has been used in a multi-disciplinary course at the University of Chicago involving computer science, linguistics and psychology students, allowing them to "speak the same language" for the course [30]. Multiple researchers can collaborate on annotations and researchers at smaller labs without such sophisticated capabilities can also make use of the datasets. Computational aspects of the work include distributed acoustic analysis using Praat, statistical analysis using R, and matrix computations using Matlab and Octave.

The SIDGrid development team has also expanded in other areas, developing google gadgets that interface to the TeraGrid. Examples developed by this group and others can be found at igoogle.com, by clicking on "add gadgets" and searching for TeraGrid. Developers of the future may start from an igoogle page and just add gadgets to develop a gateway. The team's work includes inter-gadget communication, single sign-on across gadgets, data management and ensemble support using Swift [31]. Many papers have been presented on this work at TeraGrid conferences.

GridChem [23]. (Pamidhantam, Marru) The GridChem gateway has taken a group already comfortable with scientific computing - computational chemists - and made them more effective by eliminating tedious tasks such as tracking large numbers of parametric workflows and chaining jobs from different software packages together. The gateway is leading in other areas such as scheduling and software deployment.

UltraScan [32]. (Demeler, Brooks, Marru) UltraScan is analysis and modeling software used by biochemists, biophysicists, and material scientists developed by Dr. Borries Demeler and his group at UT Health Science Center San Antonio (UTHSCSA) for the analysis of analytical ultracentrifugation experiments. The UltraScan gateway allows thousands of users to make use of the ultracentrifuge at UTHSCSA and the analysis software or upload results from a local instrument and use just the software (Figure 2). The production gateway is used by researchers throughout the world and also makes use of computational resources throughout the world in addition to TeraGrid (now XSEDE) resources. TeraGrid/XSEDE supporters have been able to apply fault-tolerance techniques from other gateways in working with the UltraScan team to achieve their goals.

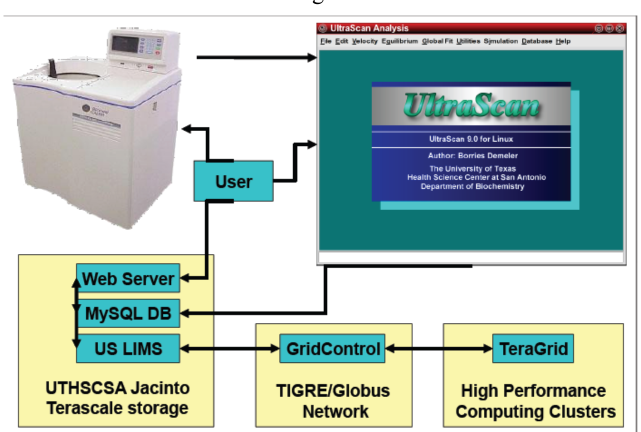

Figure 2 UltraScan workflow. Image courtesy Emre Brooks, UTHSCSA

Ocean Land Atmosphere Simulation (OLAS) [33]. (Mattocks, Marru)This project will provide gateway-based hurricane and storm surge modeling and is also intended for classroom use. Unidata LDM data ingesting tools are deployed to fetch real time data streams from the National Hurricane Center and other NOAA sources. ADCIRC tidal simulations are then run. The workflow has been demonstrated on the 268,000 node NC Grid for Hurricane Ike simulations and will be a nice test of on-demand computing capabilities in XSEDE.

Earth System Modeling Framework Web Services [34]. (DeLuca, You), This work was initiated through a helpdesk ticket for a group at NOAA. The goal is a unique chaining of a hydrological model running on a desktop and a climate modeling code (CCSM/CESM) running on a supercomputer. Early work has involved security personnel in addition to gateway developers to assess how best to enable this in the XSEDE environment. Lessons learned will be relevant for other applications of this type and perhaps those involving computational steering.

GISolve [35]. (Wang, Liu) Geographic Information Science (GIScience), an interdisciplinary field involving geography and other social sciences, computer and computational sciences, geodesy, information sciences, and statistics to study generic issues about the development and use of geographic information systems (GIS) technologies. Since GIS is used as a framework by many communities, the GISolve gateway has been an avenue for broadening participation in the TeraGrid. Communities using GISolve include those interested in plant biology, food supply, wildlife migration, land use and public health. A spatially-explicit agent-based model has been used to discover the spatiotemporal patterns of elk and their driving processes within the Greater Yellowstone Ecosystem.

A hands-on tutorial was conducted at the USGS Center for Excellence for Geospatial Information Science (CEGIS) to study hydrological data involving environmental contaminants from agricultural process and a data transfer service initiated between the USGS National Map and TeraGrid. GISolve was also featured on the cover of the Proceedings of the National Academy of Sciences (Figure 3, Reference 26).

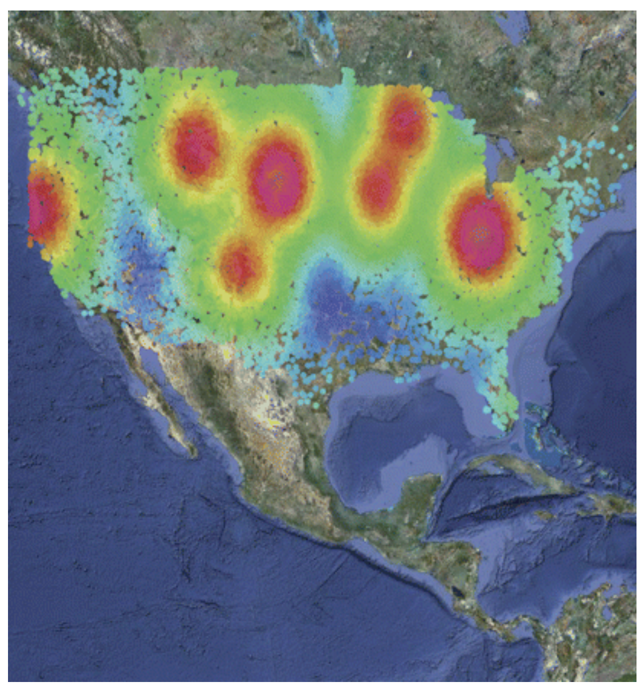

Figure 3 GISolve on the cover of PNAS, April 5, 2001. Image courtesy Shaowen Wang, NCSA.

Community Climate System Model (CCSM), Community Earth System Model (CESM), Earth System Grid (ESG) [36, 37]. (Middleton, DeLuca, Song, Zhao) The long term vision for this

Gateway is a semantically enabled environment that includes modeling, simulated and observed data holdings, and visualization and analysis for climate as well as related domains. Purdue has developed a mature gateway that provides experts and non-experts with access to the NCAR-developed CCSM (now CESM) code. ESG provides integrated data management, access, analysis, and visualization on a grid that is global in scale and serves more than 20,000 users, along with the World Climate Research Program (WCRP) and the Intergovernmental Panel on Climate Change (IPCC). ESG's structure requires the metadata necessary for reproducibility and validation, an increasingly important component of climate simulations because of their impact on policy. CESM results from the Purdue portal will now be stored in ESG. "Developing an Integrated End-to-end TeraGrid Climate Modeling Environment", submitted to the Gateway track at TeraGrid 11 won best paper in that track and was co-authored by CCSM/CESM/ESG gateway team.

Purdue's CCSM gateway [38] gives students and researchers hands-on access to a world-class, sophisticated climate code running on the TeraGrid. The portal was used in a graduate course in 2009, "Models in Climate Change Science and Policy" (Huber). The goal of the course was to better educate students in the use and misuse of modeling across disciplines, so that they will be better builders and consumers of integrated, complex models in their own careers. Students generated policy recommendations based on their own analysis of output from a suite of scientific, economic, and political models of climate change impacts (Figure 4).

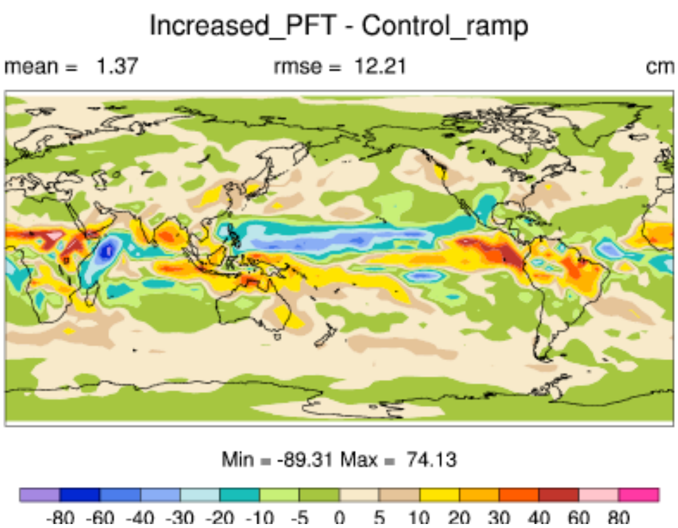

Figure 4 Students increase forestation hoping to increase precipitation and decrease temperature. Image courtesy Carol Song, Purdue.

Cyberinfrastructure for End-to-End Environmental Exploration Portal (C4E4 [39]*).* (Zhao) This gateway provides online model simulations of SWAT (Soil Water Assessment Tool [40]) and HRLDAS (High Resolution Land Data Assimilation System [41]). Figure 5 shows a SWAT view of an Indiana watershed. An auto calibration test via the gateway resulted in performance improvements of over a factor of three over the conventional approach. Future work for this gateway, includes incorporation of remote sensing from sources such as the National Climactic Data Center, the USGS, the United States Historical Climatology Network and the Climate Prediction Center. The incorporation of sensor streams will be a very useful prototype for other gateways.

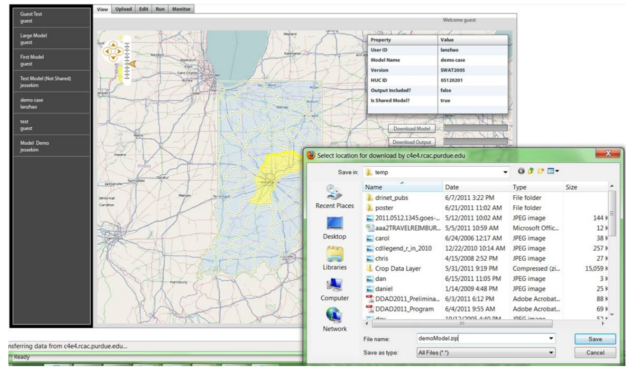

Figure 5 Soil Water Assessment Tool (SWAT) view of an Indiana watershed. Image courtesy Lan Zhao, Purdue.

RENCI Bioportal [10]. (Reed, McGee) The RENCI Bioportal was unique among TeraGrid gateways in several ways. First, it provided access to far more applications than any other gateway - over 140 computational tools and 300 GB of biological data within a collaborative workspace. The RENCI team developed a Pacman cache to streamline application installation and included web service interfaces to all services. In 2006 the RENCI Bioportal (Reed, McGee) won the *GridToday* award for Most Innovative Grid Implementation in Life Sciences.

The team also developed a more general technique for listing software programmatically. Gateways with large numbers of software offerings could make their listings available to TeraGrid in a standard fashion. The Bioportal supported Condor glide-ins and the use of OSG as well as TeraGrid and provided expertise to other gateway developers interested in the same.

Unfortunately the gateway is no longer operational, due to the lack of a sustainable funding stream. This can be a problem for many gateways and is the subject of an NSF study entitled "Opening Science Gateways to Future Success: The Challenges of Gateway Sustainability" [42]. We often observe a gateway development cycle where in a 3-5 year project requirements are gathered, a prototype is built, friendly scientists are involved, improvements are made after a beta release and just as the gateway is ready for production is beginning to attract more scientists, the project comes to an end. This cycle is often wasteful and can leave scientists mistrustful of cyberinfrastructure developments.

BioDrugScreen [43]. (Meroueh) This gateway (Figure 6) provides data and scoring function tools for both pre-docked and on-demand docking of drug-like molecules with protein targets from the Protein Data Bank. It also allows drug researchers to develop their own scoring functions of calculating how well a drug will interact with a protein. It's a fine example of a gateway that pulls from a database and performs more intensive computations than one could achieve on a desktop.

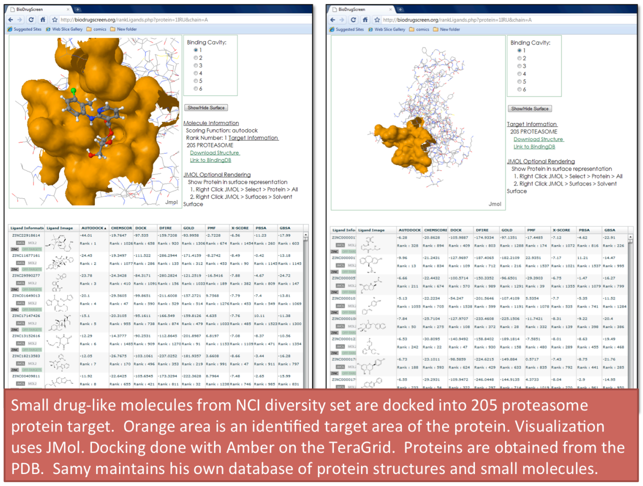

Figure 6 Biodrugscreen gateway snapshot. Image courtesy Marlon Pierce, Indiana U.

iData. (Song, Zhao) iData is a general purpose gateway developed at Purdue to enable individual researchers to self publish and share their datasets. The project supports both file-based and table-based data. It provides metadata support, condition and key word queries and data upload, download, modification, exploration, access control and usage reports. The enabling of individuals is a very powerful model for the gateway program. The first version of iData has been completed, delivering a real-time data collection of water quality parameters for Massie Creek, Ohio for the Drought Research Initiative Network (driNET) project and includes a mentoring focus involving undergraduates at minority-serving institutions.

Expressed Tag Sequence (EST) [44]. (Pallickara) The increased availability of next generation sequencing devices is producing very large volumes of data. Successful gateways in this area can further significant scientific advances. The EST gateway creates pipelines of applications running on TeraGrid compute systems available to analyze this data. Applications include RepeatMasker, PACE and CAP3 and the pipelines can be quite complex to manage, with run times ranging from milliseconds to days. Benchmark tests have been completed on up to 2M sequences.

6. CURRENT INFRASTRUCTURE HIGHLIGHTS

Quarry. (Lowe) Indiana University provides a unique allocable resource providing virtual hosting of RPM-based Linux distributions. Quarry is a set of Dell AMD systems geographically distributed to provide failover capability, is professionally managed and can be used by researchers for a variety of purposes. The system is particularly popular for gateway hosting because of its network connectivity. Pre-built virtual machines are available with software useful for gateway building, including OGCE and SimpleGrid. CI Tutor asynchronous training modules will also be available for both OGCE and SimpleGrid.

Apache Software Foundation. (Pierce, Marru) The IU Science Gateway group has initiated an Apache Software Foundation incubator project, Apache Airavata [45], in an effort to make science gateway software not just open source but open community. Airavata concentrates on software needed for distributed job and workflow management. The Apache Rave project [46], provides general-purpose software for building the Web front ends for collaborative gateways. The team hopes to demonstrate through these Apache projects that the Apache Software Foundation will provide a neutral arena for science gateway (and more generally, cyberinfrastructure) software that is controlled by the community rather than a specific institution. If successful, this will eliminate redundant development, properly distribute credit, increase sustainability, and increase the integration of gateway software with the broader Apache community.

Helpdesk. (Liu) With the increasing popularity of science gateways, the helpdesk team serves an important role as the first contact many researchers have with gateway staff. Helpdesk staff need to be able to quickly understand complex questions from those not familiar with the TeraGrid or the gateway program and provide accurate advice. They must have a thorough understanding of the gateway program and also be alert to when a helpdesk request ought to become a full-fledged project with dedicated staff.

Career development. The gateway program has developed a cadre of talent. Employers recognize the skill required to develop a successful gateway and our staff members are often sought after in the job market. Students working with gateway staff at one TeraGrid site have been hired on at other sites. Developers looking for work who inquire on the gateway list have received multiple job offers. We have developed a tremendous pool of talent, with the ideal skills to translate between science needs and CI implementation.

Contributions to stability of the grid environment. (Marru) The gateway team has participated in debugging efforts lasting months. When problems were encountered during periods of intensive use (for example the LEAD student Weather Challenge), the group assembled afterward to analyze what improvements needed to be made so that future period of intensive use could be handled in a more routine fashion. Gateway developers spent large amounts of time developing tests to simulate periods of intense load and working with large teams of resource administrators and grid software developers to understand the root causes of the failures. Testing procedures have been made available via SourceForge and inca [47] tests have been developed to simulate gateway load so that future tests can be conducted without impacting a science team. Gateway staff were also heavily involved in GRAM5 testing. Staff developed a test suite that included a range of job sizes. They discovered problems with failed jobs not being released and the problems were corrected before the software was put into production. Debugging experiences have also been shared with the Open Science Grid.

Attribute-based authentication and accounting [48]. (Basney) Because the use of community accounts anonymizes access to the TeraGrid it also removes the ability to programmatically count the number of end users using TeraGrid and now XSEDE systems. Attribute-based authentication, allows gateways to send additional information in their certificates that uniquely identify the end users on a given gateway. When this approach is in wide use, gateway developers will be able to make use of per-user usage information stored in TeraGrid/XSEDE databases. Tools for retrieving and analyzing this information have also developed through the gateway program. Several gateway developers shared early implementation experiences and wrote these up in papers in an effort to assist others performing the same upgrades.

Software listing. Efforts are underway to include software available through gateways in the comprehensive listing of software available on all XSEDE platforms. This will give users more choice when considering the tools they have available to conduct their science using a given piece of software.

7. LESSONS LEARNED AND CONCLUSIONS

The gateway program began with detailed interviews with 10 established cyberinfrastructure projects and investigated what changes would be required to harness NSF's high-end computing facilities and use them for analysis through these environments. While a requirements analysis is many times the way to begin a new program, I believe this approach was particularly effective for several reasons. First, it focused on existing, mature projects that understood their needs to a certain degree. Second, the projects continued to be involved in an ongoing way as solutions were developed. Third, we went into the interviews asking about requirements in defined areas rather than asking open-ended questions and so came out with fairly concrete recommendations.

This approach laid the groundwork for the program going forward. While some requirements turned out to be distractions, others led to capabilities that became fundamental components of the gateway program. The program could have moved along more quickly if we had more quickly put an end to development and policy where we did not have a large user base, however it wasn't always possible to tell at the outset whether a requirement would turn out to be fundamental or spurious.

It is true that the choice of initial prototype projects and development of infrastructure that met their needs meant that other needs were not being met and that a certain class of applications, for example those utilizing web services, was therefore finding TeraGrid less relevant. This perpetuates itself as the program attracts more projects who see what they need and less who don't.

While the 5-year plans for the first gateways hampered our ability to support new projects, that transition when it finally happened was a turning point in the program. While TeraGrid staff could concentrate on issues associated with the application of high-end cyberinfrastructure to the back end of gateways, front-end gateway developers remained with the science communities and were in the best position to remain responsive to their users. Because of their experience with large numbers of projects, TeraGrid staff were often in a position to make recommendations, however they were not building gateways for others.

It is also important to document progress, particularly in the area of policy agreements. If not, contentious topics are often revisited and time wasted.

Once common infrastructure was identified, clear and simple documentation describing how to use it also made the program more scalable. While the TeraGrid's gateway program focused on back-end integration, it is true that the team saw many front-end interfaces and was often asked to provide recommendations in such areas. Build-a-gateway in a day tutorials, downloadable software for simple gateway building and later virtual machines pre-installed with gateway software and asynchronous CI Tutor modules were all important contributions and demonstrated the necessary underpinnings of a functional science gateway. It may be that gateway building can be even further simplified through the use of igoogle and pre-defined gadgets. We did not keep statistics on those who inquired about the program and followed through and built a gateway vs. those who inquired and did not move ahead. This would be an interesting data point as well.

Other groups have also developed very successful gateway-building software. HUBzero [49], developed by the originators of nanoHUB has been extraordinarily successful and is used by gateways in many different domains including healthcare, manufacturing, volcano research, pharmaceutical product development and STEM (science, technology, engineering and math) education, as well as nanoscience. We've seen some researchers use frameworks such as Drupal [50] and Django [51] quite successfully for front-end development.

Gateways can be quite full-featured or very simple. They can be developed and run with a fraction of a person or with a large team. But they are not a solution to all problems. The most important component is an initial assessment of and continual contact with the user community. User communities can be quite varied in the levels of cyberinfrastructure they wish to be exposed to in a gateway. Some will want fine-grained control of a code while others want results without knowing or caring where a simulation is run.

Ease of use, functionality and reliability are three important considerations. While web pages developed in academic institutions were on the leading edge in the early days of the web, this is not true today. Science gateway users expect a science web page to be as straightforward, full featured, reliable, community-building and easy to use as commercial websites such as Google, Amazon or Facebook. nanoHUB [52] is an example of a gateway that provides such an environment, as evidenced by their user community of over 100,000.

But the answer to what makes a successful gateway is more complex. We have seen successful gateways without a streamlined user interface that meet a well-defined need, providing turn around in a day rather than a month. Gateways that meet a niche need and deliver results reliably and quickly will grow in popularity despite a lesser user interface.

Having exemplar projects and staff within the program was very helpful for showing others that "it could be done" successfully and how to do "it". In this context, "it" could mean many things - writing an allocation request, building an 800-person user community, developing a robust operating and accounting system or implementing attribute-based authentication. A successful example speaks many times louder than pages of documentation.

Including a diverse set of projects, both initially and going forward was very beneficial and allowed different domain groups to interact in ways they had not previously. Identifying components of each project that could impact others furthered the program as well.

Project teleconferences were designed not only to check the status of staff member's work and allow them to highlight accomplishments, but also to bring in outside speakers. The calls became a forum of sorts, attracting many attendees who were not funded on the TeraGrid project. TeraGrid staff could learn about developments elsewhere and outsiders could learn how our program was being run. Both of these were extremely beneficial.

When tackling large project-wide efforts such upgrade to GRAM5 as the default implementation on all systems, development of common approaches to securing community accounts at all sites, or the requirement that all gateways implement attribute-based authentication, it is best to tackle one big project only and follow it closely to its conclusion, continually pressing to make sure the work moves ahead. By tackling several of these at once, the project duration stretched out significantly and success became much more difficult to achieve. In a distributed project where there is much vying for a staff member's attention, these delays can undermine credibility the next time there is an extended goal to achieve.

While the consumption of CPU time by gateways is still a small percentage (1.5% in 2011) of that used by command line users, their impact has been high. They democratize computational science by providing access to high-end resources to all. Researchers at over 1000 institutions make use of CIPRES where they can generate publication-quality results overnight rather than the month it can take for some calculations on a desktop. This sort of "time to science" improvement can fundamentally change the way researchers consider problems, regardless of their location The SIDGrid gateway provides access to difficult and expensive to assemble datasets that can include fMRI image data, heart rate, eye tracking, video and annotation capabilities underpinned by high performance computing where it's needed. This assemblage has contributed to an interdisciplinary approach in fields such as computational linguistics.

In 2011, 99% of gateway CPU use has been on clusters rather than shared memory platforms, but that is likely a reflection of the mix of available resources rather than gateway suitability. In truth, any platform (cluster, shared memory, GPU machine, Condor flock) can back-end a gateway successfully. It all depends where a gateway's applications run well. We have seen usage on all types of systems at different points in time.

8. FUTURE WORK

While this paper has focused primarily on the development and operation of the gateway program, an additional paper of interest would be a survey of gateway development efforts. What technologies are used to build gateways, how do different gateways conduct their job submission and data transfer, how do they do accounting, how do they attract a user community? While we have collected some of this data through our work in the TeraGrid program (Martin, http://teragridforum.org/mediawiki/index.php?title=Science_Gateway_Use_Cases), a full survey paper would be very useful.

The TeraGrid and now XSEDE science gateways work tends to be very operationally oriented. The program's responsibility is to ensure that community designers can make use of the high-end infrastructure without undue amounts of effort. This is no small task and continues to remain challenging. Policy agreement that leads to simplification for developers can be hard fought and need to continue to be maintained as service providers change. Grid software infrastructure on XSEDE is diversifying. Reliability and ease of use will remain challenges.

Investigation of clouds as a back end resource could be a nice model for some gateways because of the lack of queue waits, the expandable nature of clouds, the individual control, etc. Depending on the nature of the algorithms being solved, this could be a nice match for some gateways. An allocated cloud implementation for researchers to experiment with would be beneficial. It may be that FutureGrid [53] can help researchers explore the suitability of this computing model.

Finally, sustainability is a major challenge for gateways. Gateways are tools and sometimes the only tools that can be used to tackle transformative research problems. Sometimes building them and maintaining them is not a transformative activity in and of itself however and therefore can be difficult to fund as research projects. The Office of Cyberinfrastructure has been very forward-thinking on sustainability issues, but there are many challenges that remain.

One component of sustainability is the identification of projects that ought to be funded in a continuing way. Which gateways are having the most impact? What makes an effective gateway? How ought projects to be evaluated? What factors combine to engender success? Some recommendations that may lead to sustainable gateways have been considered in a recent and ongoing study [42] and include approaches such as staged funding, incubator-like support and community support for gateway developers.

9. ACKNOWLEDGMENTS

There are a great many individuals who were foundational participants in the science gateway program. I apologize for important names I have inadvertently left off this list. It has been a tremendous honor working with Jay Alameda, Jim Basney, Pete Beckman, Julian Bunn, Charlie Catlett, Steve Clark, John Cobb, Marcus Christie, Cecelia DeLuca, Diana Diehl, Rion Dooley, Kelvin Droegemeier, Dennis Gannon, Kevin Gamiel, Victor Hazelwood, Ivan Judson, Dan Katz, Gerhard Klimeck, Yan Liu, Rich Loft, Mike Lowe, Vickie Lynch, Jim Marsteller, Stu Martin, Mark Miller, John McGee, Don Middleton, Steve Mock, Sudhakar Pamidighantam, Wayne Pfeiffer, Marlon Pierce, Lavanya Ramakrishnan, Jason Reilly, Eric Roberts, Anurag Shankar, Ashish Sharma, Eric Shook, Raminder Singh, Preston Smith, Rick Stevens, Carol Song, Christopher Thompson, Shaowen Wang, Von Welch, Roy Williams, Matthew Woitaszek, Lan Zhao.

10. REFERENCES

[1] http://en.wikipedia.org/wiki/Mosaic_%28web_browser%29, retrieved 10/25/11.

[2] http://data.worldbank.org/data-catalog/world-development-indicators?cid=GPD_WDI, data published at http://www.google.com/publicdata, retrieved 10/25/11.

[3] Miller, M.A.; Pfeiffer, W.; Schwartz, T.; , "Creating the CIPRES Science Gateway for inference of large phylogenetic trees," *Gateway Computing Environments Workshop (GCE), 2010* , vol., no., pp.1-8, 14-14 Nov. 2010. doi: 10.1109/GCE.2010.5676129. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5676129&isnumber=5676117

[4] http://www.nsf.gov/pubs/1999/nsf99167/nsf99167.htm, retrieved 10/25/11.

[5] B. Joy and K. Kennedy, editors. 1999. Information Technology Research: Investing in Our Future. President's Information Technology Advisory Committee, http://www.nitrd.gov/pitac/report/index.html, retrieved 10/25/11.

[6] Droegemeier, K.K., et al. Linked Environments for Atmospheric Discovery (LEAD): A CyberInfrastructure for Mesoscale Meteorology Research and Education. in *20th Conference on Interactive Information Processing Systems for Meteorology, Oceanography, and Hydrology*. 2004. Seattle, WA.

[7] Szalay, A. S. The National Virtual Observatory in *Astronomical Data Analysis Software and Systems X, ASP Conference Proceedings*, Vol. 238. Edited by F. R. Harnden, Jr., Francis A. Primini, and Harry E. Payne. San Francisco: Astronomical Society of the Pacific, ISSN: 1080-7926, 2001., p.3

[8] Klimeck, G.; McLennan, M.; Brophy, S.P.; Adams, G.B.; Lundstrom, M.S.; , "nanoHUB.org: Advancing Education

and Research in Nanotechnology," *Computing in Science & Engineering* , vol.10, no.5, pp.17-23, Sept.-Oct. 2008
doi: 10.1109/MCSE.2008.120
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4604500&isnumber=4604492

[9] The National Microbial Pathogen Database Resource (NMPDR): A genomics platform based on subsystem annotation. McNeil LK, Reich C, Aziz RK, Bartels D, Cohoon M, Disz T, Edwards RA, Gerdes SY, Hwang K, Kubal M, Margaryan GR, Meyer F, Mihalo W, Olsen GJ, Olson R, Osterman AL, Paarmann D, Paczian T, Parrello B, Pusch GD, Rodionov DA, Shi X, Vassieva O, Vonstein V, Zagnitko OP, Xia F, Zinner J, Overbeek R, Stevens R. Nucleic Acids Res. 2007 Jan;35(Database issue):D347-53. [Epub 2006 Dec 1] PubMedID 17145713

[10] Blatecky A, Gamiel K, Ramakrishnan L, Reed D, Reed M. Building the bioscience gateway. *Proceedings of the Workshop on Science Gateways: Common Community Interfaces to Grid Resources at GGF-14*, June 2005. Global Grid Forum: Lemont, IL, 2005.

[11] Cobb, J. W., Geist, A., Kohl, J. A., Miller, S. D., Peterson, P. F., Pike, G. G., Reuter, M. A., Swain, T., Vazhkudai, S. S. and Vijayakumar, N. N. (2007), The Neutron Science TeraGrid Gateway: a TeraGrid science gateway to support the Spallation Neutron Source. *Concurrency and Computation: Practice and Experience*, 19: 809–826. doi: 10.1002/cpe.1102

[12] F. van Lingen, J. Bunn, I. Legrand, H. Newman, C. Steenberg, M. Thomas, P. Avery, D. Bourilkov, R. Cavanaugh, L. Chitnis, M. Kulkarni, J. Uk In, A. Anjum, T. Azim "*Grid Enabled Analysis: Architecture, Prototype and Status*" in proceedings of Computing for High Energy Physics (CHEP) Interlaken, Switzerland September 2004. (see also: http://ultralight.caltech.edu/gaeweb/portal)

[13] Pordes, R. (2007). The open science grid. Journal of physics. Conference series, 78, 012057-.

[14] Thomas, M.; Mock, S.; Dahan, M.; Mueller, K.; Sutton, D.; Boisseau, J.R.; , "The GridPort toolkit: a system for building Grid portals," *High Performance Distributed Computing, 2001. Proceedings. 10th IEEE International Symposium on* , vol., no., pp.216-227, 2001
doi: 10.1109/HPDC.2001.945191
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=945191&isnumber=20446.

[15] van Lingen, F.; Steenberg, C.; Thomas, M.; Anjum, A.; Azim, T.; Khan, F.; Newman, H.; Ali, A.; Bunn, J.; Legrand, I.; , "The Clarens Web service framework for distributed scientific analysis in grid projects," *Parallel Processing, 2005. ICPP 2005 Workshops. International Conference Workshops on* , vol., no., pp. 45- 52, 14-17 June 2005
doi: 10.1109/ICPPW.2005.82
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1488676&isnumber=32020

[16] Sumalatha Adabala, Vineet Chadha, Puneet Chawla, Renato Figueiredo, José Fortes, Ivan Krsul, Andrea Matsunaga, Mauricio Tsugawa, Jian Zhang, Ming Zhao, Liping Zhu, Xiaomin Zhu, From virtualized resources to virtual computing grids: the In-VIGO system, Future Generation Computer Systems, Volume 21, Issue 6, June 2005, Pages 896-909, ISSN 0167-739X, 10.1016/j.future.2003.12.021. (http://www.sciencedirect.com/science/article/pii/S0167739X03002899)

[17] Alameda, J. (2007). The Open Grid Computing Environments collaboration: portlets and services for science gateways. *Concurrency and Computation: Practice and Experience*, 19(6), 921-.

[18] Novotny, J. (2004). GridSphere: a portal framework for building collaborations. *Concurrency and Computation: Practice and Experience*, 16(5), 503-.

[19] http://teragridforum.org/mediawiki/index.php?title=TeraGrid_Science_Gateways_Primer, retrieved 10/25/11.

[20] Beckman, P. (2007). SPRUCE: A system for supporting urgent high-performance computing. In *Grid-Based Problem Solving Environments.* (pp. 295-).

[21] WilkinsDiehr, N, & Wilkins-Diehr. (2007). Special issue: Science gateways—Common community interfaces to grid resources. *Concurrency and Computation: Practice and Experience*, 19(6), 743-.

[22] Wenjun Wu; Uram, T.D.; Wilde, M.; Hereld, M.; Papka, M.E.; , "A Web 2.0-Based Scientific Application Framework," *Web Services (ICWS), 2010 IEEE International Conference on* , vol., no., pp.642-643, 5-10 July 2010
doi: 10.1109/ICWS.2010.107
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5552726&isnumber=5552704

[23] Wilkins-Diehr, N., Gannon D., Klimeck, G., Oster, S., Pamidighantam, S. 2008. "TeraGrid Science Gateways and Their Impact on Science", *IEEE Computer*, volume 41, number 11, November 2008.

[24] Allan, R. 2009. *Virtual Research Environments: From Portals to Science Gateways*. Chandos Publishing, Oxford.

[25] Wilkins-Diehr, N., Baru, C., Gannon, D., Keahey, K., McGee, J., Pierce, M., Wolski, R., Wu, W. 2011. "Science Gateways: Harnessing Clouds and Software Services for Science" in *Cloud Computing and Software Services: Theory and Techniques*, 17-46. Edited by Syed A. Ahson and Mohammad Ilyas. CRC Press, Boca Raton, FL.

[26] "The Emergence of Spatial Cyberinfrastructure" Dawn J. Wright, Shaowen Wang. doi:10.1073/pnas.1103051108 *PNAS April 5, 2011 vol. 108 no. 14 5488-5491.*

[27] Shaowen Wang, Yan Liu, Nancy Wilkins-Diehr, Stuart Martin, SimpleGrid toolkit: Enabling geosciences gateways to cyberinfrastructure, *Computers & Geosciences*, Volume 35, Issue 12, December 2009, Pages 2283-2294, ISSN 0098-3004, 10.1016/j.cageo.2009.05.002. (http://www.sciencedirect.com/science/article/pii/S0098300409002210)

[28] Woitaszek, M. (2009). AMP: A Science-driven Web-based Application for the TeraGrid. In *Proceedings of the 5th Grid Computing Environments Workshop - GCE '09*. (pp. 1-).ACM.

[29] Bertenthal, B., Grossman, R., Hanley, D., Hereld, M., Kenny, S., Levow, G., Papka, M. E., Porges, S., Rajavenkateshwaran, K., Stevens, R., Uram, T., and Wu, W. (2007) "Social Informatics Data Grid,", *Third International Conference on e-Social Science.* October 7-9, 2007, Ann Arbor, Michigan

[30] Gina-Anne Levow. 2008. Studying discourse and dialogue with SIDGrid. In *Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics* (TeachCL '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 106-113.

[31] Yong Zhao; Hategan, M.; Clifford, B.; Foster, I.; von Laszewski, G.; Nefedova, V.; Raicu, I.; Stef-Praun, T.; Wilde, M.; , "Swift: Fast, Reliable, Loosely Coupled Parallel Computation," Services, 2007 IEEE Congress on , vol., no., pp.199-206, 9-13 July 2007. doi: 10.1109/SERVICES.2007.63. URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4278797&isnumber=4278749

[32] Demeler, B. 2005. UltraScan: A comprehensive data analysis software package for analytical ultracentrifugation experiments. In *Modern analytical ultracentrifugation: Techniques and methods* (eds. D.J. Scott et al.), pp. 210–230. Royal Society of Chemistry, Cambridge, UK.

[33] Walko,R.L.,Avissar,R. OLAM:Ocean-Land-AtmosphereModel-ModelInputParameters- Version 3.0. Tech. Rep., Duke University , November, 2008.

[34] Collins, N., Theurich, G., DeLuca, C., Suarez, M., Trayanov, A., Balaji, V., Li., P., Yang, W., Hill, C., and da Silva, A. (2005). Design and implementation of components in the Earth system modeling framework, *International Journal of High Performance Computing Applications*, **19**(3): 341–350.

[35] Shaowen Wang; Armstrong, M.P.; Jun Ni; Yan Liu; , "GISolve: a grid-based problem solving environment for computationally intensive geographic information analysis," *Challenges of Large Applications in Distributed Environments, 2005. CLADE 2005. Proceedings* , vol., no., pp. 3- 12, 24 July 2005
doi: 10.1109/CLADE.2005.1520892
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1520892&isnumber=32532

[36] Collins, W.D. and Coauthors, 2006a: The Community Climate System Model version 3 (CCSM3). J. Climate, 19, 2122–2143.

[37] Bernholdt, D., Bharathi, S., Brown, D., Chanchio, K., Chen, M., Chervenak, A., Cinquini, L., Drach, B., Foster, I., Fox, P., Garcia, J., Kesselman, C., Markel, R., Middleton, D., Nefedova, V., Pouchard, L., Shoshani, A., Sim, A., Strand, G. and Williams, D. The Earth System Grid: Supporting the Next Generation of Climate Modeling Research. *Proceedings of the IEEE*, *93* (3). 485-495. 2005.

[38] A. Basumallik, L. Zhao, C. X. Song, R. L. Sriver, and M. Huber. A Community Climate System Modeling Portal for the TeraGrid. In *TeraGrid07 Conference*, Madison, WI, June 2007.

[39] Govindaraju, R., B. Engel, D. Ebert, B. Fossum, M. Huber, C.T. Jafvert, S. Kumar, V. Merwade, D. Niyogi, L Oliver, S. Prabha- kar, G. Rochon, C. Song, and L. Zhao, 2009. A Vision of Cyber- infrastructure for End-to-End Environmental Explorations (C4E4). *Journal of Hydrologic Engineering* 14(1):53-64.

[40] Arnold, J G. (2005). SWAT2000: Current Capabilities and Research Opportunities in *Applied Watershed Modelling.HydrologicalPprocesses*, 19(3), 563-.

[41] Chen, F., and J. Dudhia, K. W. Manning, D. N. Yates, M. A. LeMone, S. B. Trier, R. Cuenca, and D. Niyogi, 2004: Development of a High Resolution Land Data Assimilation System (HRLDAS). Pre- prints, 16th Conf. on Numerical Weather Prediction, Seattle, WA, Amer. Meteor. Soc., CD-ROM, 22.3.

[42] Wilkins-Diehr, N., and Lawrence, K. A. 2011. "Opening Science Gateways to Future Success: The Challenges of Gateway Sustainability." Gateway Computing Environments Workshop (GCE), 2010 (Supercomputing '10 in New Orleans, LA), 14 Nov. 2010. IEEE Computer Society (Xplore Digital Library).

[43] Li, L., Bum-Erdene, K., Baenziger, P. H., Rosen, J. J., Hemmert, J. R., Nellis, J. A., et al. (2010). BioDrugScreen: a computational drug design resource for ranking molecules docked to the human proteome. *Nucleic Acids Research*, 38, D765–D773.

[44] S. L. Pallickara, M. Pierce, Q. Dong, and C. Kong, "Enabling Large Scale Scientific Computations for Expressed Sequence Tag Sequencing over Grid and Cloud Computing Clusters", *PPAM 2009 Eighth International Conference on Parallel Processing and Applied Mathematics.* Wroclaw, Poland, September 13-16, 2009.

[45] http://incubator.apache.org/airavata/

[46] http://incubator.apache.org/rave/

[47] Smallen, S.; Olschanowsky, C.; Ericson, K.; Beckman, P.; Schopf, J.M.; , "The Inca Test Harness and Reporting Framework," *Supercomputing, 2004. Proceedings of the ACM/IEEE SC2004 Conference* , vol., no., pp. 55, 06-12 Nov. 2004. doi: 10.1109/SC.2004.56.
URL: http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=1392985&isnumber=30315

[48] Barton T, Basney J, Freeman T, Scavo T, Siebenlist F, Welch V, Ananthakrishnan R, Baker B, Keahey K. Identity federation and attribute-based authorization through the globus toolkit, Shibboleth, Gridshib, and MyProxy. *Fifth Annual PKI R&D Workshop*, Gaitherburg, U.S.A., April 2006. Available from http://middleware.internet2.edu/pki06/proceedings/welch-idfederation.pdf.

[49] M. McLennan, R. Kennell, "HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering," Computing in Science and Engineering, 12(2), pp. 48-52, March/April, 2010.

[50] http://www.drupal.org

[51] http://www.djangoproject.com

[52] G. Klimeck et al., "nanoHUB.org: Advancing Education and Research in Nanotechnology," *Computing in Science and Engingeering*, Sept./Oct. 2008, pp. 17-23.

[53] Geoffrey Fox. FutureGrid Platform FGPlatform: Rationale and Possible Directions (White Paper). 2010 Available from: http://grids.ucs.indiana.edu/ptliupages/publications/FGPlatform.docx.