The Next Generation Research Grid: A Path Forward
Final Report

August 2008

Timothy L. Killeen, Chair
National Center for Atmospheric Research

Roberta Balstad, Associate Chair
Columbia University

Arthur Bland
Oak Ridge National Laboratory

Roscoe Giles
Boston University

Myron Gutmann
University of Michigan

Gwendolyn Huntoon
Pittsburgh Supercomputing Center

Gerhard Klimeck
Purdue University

Paul Messina
Caltech and Argonne National Laboratory (retired)

B. Montgomery Pettitt
University of Houston

Edward Seidel
Louisiana State University

Joan-Emma Shea
University of California, Santa Barbara

Alexander Szalay
The Johns Hopkins University

Disclaimer

The TeraGrid Planning Process is supported by NSF Award # OCI-0724300 to the University of Michigan. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

Executive Summary

This is the final report from a project funded under NSF award OCI-0724300 to work with stakeholder communities to collect ideas for the next generation of the TeraGrid. TeraGrid currently uses high-speed network connections to integrate high-performance computers, data resources and tools, and experimental facilities at eleven resource provider sites around the country. To address changes that are already occurring and are anticipated to take place in high-performance computing (HPC) and computational science over the next 5-7 years, the National Science Foundation (NSF) awarded a grant to the University of Michigan's School of Information (UM-SI) to facilitate a planning process to help guide the future evolution of TeraGrid. The Principal Investigators (PI) of the planning grant convened a steering committee, representative of key stakeholder communities to help achieve the project goals. Our committee was charged to provide a report to stakeholders that identifies options for the definition, design, and implementation of the next generation of the NSF TeraGrid program. In conducting our charge, we considered the results of a series of planning workshops, hosted "town hall" meetings, solicited position papers from current TeraGrid users and other national and international stakeholders, examined relevant reports and other documents, incorporated information from the TeraGrid Evaluation Research Study, interacted individually with stakeholders, and deliberated extensively.

We strongly endorse a next phase for the TeraGrid program, which we refer to hereafter as the *Next Generation Research Grid (NGRG)*. The NGRG should have an unwavering focus on science: advancing and accelerating science and engineering progress across a broad front. The NGRG should be an **open, reliable, extensible, high-end, cyberinfrastructure** for the provision of digital services, tools, and resources in support of research and education and a leader in creating an **international and global research cyberinfrastructure**. It should provide a **balanced and interoperable system** focused on advanced computation, data storage, management, curation and mining, visualization, and networking in the context of an **agile and robust production environment**. The NGRG should have **a co-equal emphasis on human capital development** – the education, training, and support of the people who can create, utilize, support, and extend the cyberinfrastructure to enable research discovery and learning for future generations. Finally, the NGRG should evolve its technical architecture and management to provide **pathways for the integration of a wide range of cyberinfrastructure resources and new providers,** while retaining its emphasis on providing access to the most advanced high-end cyberinfrastructure facilities.

The open, agile, and robust production infrastructure that we envision for the NGRG requires a funding model to support program attributes over extended time periods; a strategic plan and scheduled reassessment of the plan at set intervals; and a transparent governance structure and management plan that includes multiple avenues for stakeholder participation. To ensure that the NGRG has the stability, direction, leadership, and community support that will be necessary to its success and to its ability

to remain agile in the face of technological change we suggest that the NSF prepare a two-step announcement of opportunity for competitive planning grants leading ultimately to the selection of **an entity to manage the NGRG**. The agreement with the selected team would last for a minimum of five years and would include regularly scheduled assessments against pre-specified metrics and options for renewal of the agreement. In addition to the product that would result from the announcement of opportunity, it is important that mechanisms be developed for other research programs supported by NSF directorates to coordinate with the Office of Cyberinfrastructure and to make use of the NGRG as an integral part of their programs and to provide incentives for alignment.

1.0 Introduction

The National Science Foundation (NSF) has been making substantial investments in high-performance computing (HPC) for more than 20 years. It can be argued that the TeraGrid and other cyberinfrastructure investments made by the NSF have literally transformed many fields of science and engineering. A few of many potential examples are listed below.

- Science gateways, some of which support both research and education, facilitate access to computing and data resources.
- The creation of data archives and the development of tools for accessing and analyzing their contents have been critical to handle the 'data deluge' faced by researchers in a number of fields. Many of these resources are available to the global research community.
- The use of cyberinfrastructure has expanded to new areas. For example, social scientists are taking advantage of resources and tools to analyze social relationships in large online communities. Scholars in the humanities are utilizing TeraGrid's data storage resources to preserve ancient manuscripts.
- The development of human capital through educational programs and activities that span all levels has helped to develop skills and expertise and broaden participation in computing.

While it is clear that HPC generally, and TeraGrid specifically, have supported many important scientific achievements, there are barriers to use of the TeraGrid that, if lowered, would increase its benefit to research and engineering communities. These barriers include constraints to participation associated with the specialized human knowledge and skills that are currently required to both employ and support the use of distributed resources. We discuss these and other issues in greater detail later in this report as they are crucial if the next phase of TeraGrid, which we hereafter refer to as the *Next Generation Research Grid (NGRG)*, is to accomplish the vision that it is uniquely positioned to achieve. We hope that one outcome of the findings presented in this report will be a new name that reflects the vision of this open, high-end, production quality infrastructure focused on supporting compelling questions scientists cannot yet address and by research questions that will surface in the next five to seven years. In order to serve this role, the NGRG must be more than a distributed computational environment and be an entity that encompasses architecture, software, and people. In the remainder of this section, we describe the process that resulted in this report, and we articulate our vision for the NGRG.

1.1 The TeraGrid Planning Process

The TeraGrid Planning Process was funded by the National Science Foundation's (NSF) Office of Cyberinfrastructure (OCI) through a grant to the University of Michigan's

School of Information (UM-SI). Ann Zimmerman and Thomas A. Finholt served as Principal Investigator (PI) and Co-PI, respectively. The role of the UM-SI was to support and facilitate a planning process to be led by a steering committee representative of key stakeholder communities and with diverse expertise. The PIs invited Timothy L. Killeen and Roberta Balstad to serve as Chair and Associate Chair of the committee, and they accepted. Drs. Killeen and Balstad then worked with the PIs to assemble a full committee. The members of the steering committee and facilitation team are listed in Appendix A.

This final report of the steering committee reflects the consensus view of its members and is based on extensive deliberation, most of which occurred in a series of ten meetings held between mid-September 2007 and the end of January 2008.[1] Our findings have been informed by input from many individuals and groups, including current and potential TeraGrid users, computer scientists and technologists, application and software developers, current TeraGrid awardees, and other national and international grid projects. In addition to interactions with individuals, we considered the results from a series of planning workshops, hosted "town hall" meetings, solicited position papers from current TeraGrid users and other national and international stakeholders, and examined relevant reports and other documents. Many of these activities also served as a means to disseminate information about the planning process as did the web site and occasional electronic newsletters.[2] Appendix B provides more detail on the committee's information gathering and dissemination activities.

1.2 Vision for the Next Generation Research Grid

Based on the activities and input described above, the steering committee articulated the following vision for the NGRG.

> The NGRG should have an unwavering **focus on science**: advancing and accelerating science and engineering progress across a broad front. The NGRG should be an **open[3], reliable, extensible, high-end, cyberinfrastructure** for the provision of digital services, tools, and resources in support of research and education and a leader in creating an **international and global research cyberinfrastructure**. It should provide a **balanced and interoperable system** focused on advanced computation, data storage, management, curation and mining, visualization, and networking in the context of an **agile and robust production environment**. The NGRG should have **a co-equal emphasis on human capital development** – the education, training, and

[1]Dane Skow, TeraGrid Grid Infrastructure Group, participated actively as an ex-officio member of the steering committee until December 2007. Since he was not able to be present at the committee's final meetings, the findings presented in this report should not be viewed as representing his views or those of TeraGrid.

[2] The planning process web site is available at http://www.teragridfuture.org.

[3] We define "open" to mean a system that is readily available for the participation of new users with the minimum of barriers to full access, and with support from a dedicated group of expert advisors and access to detailed and up-to-date documentation,

support of the people who can create, utilize, support, and extend the cyberinfrastructure to enable research discovery and learning for future generations. Finally, the NGRG should evolve its technical architecture and management to provide **pathways for the integration of a wide range of cyberinfrastructure resources** while retaining its emphasis on providing access to the most capable cyberinfrastructure facilities.

In the remainder of this report, we provide an overview of the history of the current TeraGrid program, discuss six key and inter-related issues that underpin our vision for the NGRG, and present our findings for a path forward for the next phase of the TeraGrid.

2.0 Background and Context

It was beyond our charge to document the history of the TeraGrid project or to analyze the wider contexts under which it developed and evolved. We found, though, that it was necessary to understand this background at a general level in order to conduct our work. Therefore, in this section, we provide a brief history of the TeraGrid project, including some of the activities and events that preceded it. The information presented here was gathered from publicly available sources, including the NSF web site, NSF solicitations related to TeraGrid, and several reports.

The TeraGrid, as it currently exists, evolved from a series of NSF solicitations and Dear Colleague letters and grew out of a late 1990's focus on terascale computing and grid computing. For example, in 1999, a report by the President's Information Technology Advisory Committee (PITAC) recommended that in order for "the United States to continue as the world leader in basic research, its scientists and engineers must have access to the most powerful computers" (p. 52). At this time, terascale systems (10^{12} operations per second) were emerging as the most capable systems. The PITAC report was preceded by three NSF-sponsored workshops on terascale and petascale computing that were held during May and July 1998. The findings from these workshops are documented in a consolidated report (Reed, et al., 1998). A similar joint Department of Energy/NSF workshop was held shortly after the three NSF workshops (Langer, 1998).

2.1 Brief History of TeraGrid

What is now known as the TeraGrid began in 2001 as a partnership among four sites: Argonne National Laboratory (ANL), California Institute of Technology (Caltech), National Center for Supercomputing Applications (NCSA), and the San Diego Supercomputer Center (SDSC). This collaboration was the result of a solicitation for a Distributed Terascale Facility (DTF) issued by NSF in January 2001. The solicitation stated that more than computational capability was required to meet the needs of scientists and engineers.

> Investments in large scale research instrumentation being made in such diverse fields such as astronomy, biology, earthquake engineering, environmental science,

> geosciences, gravitational science, and high energy physics, will not yield their full returns unless corresponding investments are made in the infrastructure needed for data analysis. Terascale computing systems and large-scale scientific instruments and sensors are now routinely creating multi-terabyte data archives. All the researchers involved encounter similar problems since computed, observed, and experimental data all require data manipulation and storage, visualization, data mining and interpretation. The rapidly increasing rate at which data are being generated and the distance between its point of generation and those who need access to information contained in the data are problems that must be faced (NSF, 2001).

The DTF solicitation identified a computational grid as a means to meet these needs. The solicitation defined the grid as "the sum of networking, computing, and data storage technologies needed to create a seamless, balanced, integrated computational and collaborative environment." As noted above, the DTF was named TeraGrid and included computers capable of 11.6 teraflops, disk-storage with capacity of more than 450 terabytes of data, visualization systems, and data collections integrated via grid middleware and linked through a high-speed optical network (NRC, 2004).

In April 2002, NSF issued a *Dear Colleague Letter on the Extensible Terascale Facility (ETF) for Principal Investigators* to expand the DTF. A $35 million award was made later in 2002 to expand the capabilities of the original DTF sites and to include Pittsburgh Supercomputing Center's (PSC) LeMieux system. The partnership continued to be called the TeraGrid, but the award program was known as ETF.

In March 2003, NSF issued the *Terascale Extensions Program Solicitation NSF03-553* to further expand ETF capabilities. This resulted in three separate awards to fund the high-speed networking connections needed to share resources at Indiana University (IU), Purdue University (Purdue), Oak Ridge National Laboratory (ORNL), and Texas Advanced Computing Center (TACC). Purdue and IU partnered on a proposal and received a joint award.[4] In 2004, TeraGrid entered full production, and in 2005 NSF's newly created Office of Cyberinfrastructure extended support for TeraGrid with a $150 million set of awards for operation, user support, and enhancement of the TeraGrid facility over the next 5 years.

In June 2006, the National Center for Atmospheric Research (NCAR) became the ninth TeraGrid RP site. The tenth and eleventh sites—the Louisiana Optical Network (LONI) and the National Institute for Computational Sciences (NICS) established by the University of Tennessee and Oak Ridge National Laboratory—joined the TeraGrid in fall 2007. Figure 1 shows the locations of the current TeraGrid resource providers.

[4] Since 2005 Purdue and Indiana have had separate awards.

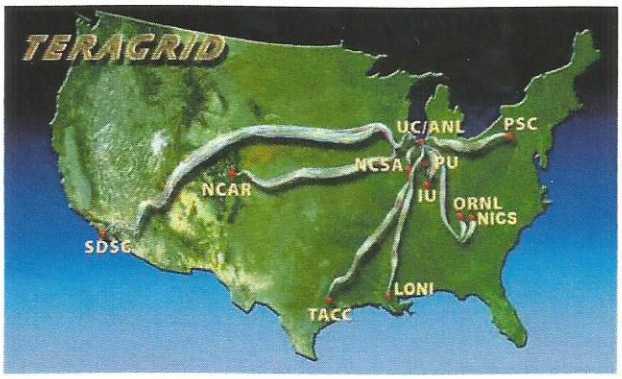

Figure 1: TeraGrid Resource Providers as of May 2008
(Image courtesy of TeraGrid)

2.2 Sources of Information

We relied on two primary sources of information in preparing the findings presented in this report: 1) results from the TeraGrid evaluation research study, and 2) information collected during the planning process. In this section, we describe each of these activities and then summarize the findings relevant to the next phase of TeraGrid.

2.2.1 TeraGrid Evaluation Research Study

In late spring 2006, the NSF awarded a grant to the University of Michigan's School of Information to conduct an external evaluation of TeraGrid. The primary goals of the evaluation were a) to provide specific information to TeraGrid managers to increase the likelihood of TeraGrid success, and b) to give NSF and policy makers general data to assist them in making strategic decisions about future directions for cyberinfrastructure. In order to accomplish these objectives, the UM-SI study assessed four aspects of the TeraGrid project:

- progress in meeting user requirements;
- impact of TeraGrid on research outcomes;
- quality and content of TeraGrid education, outreach, and training activities; and
- satisfaction among TeraGrid partners.

Researchers at UM-SI used multiple methods to collect and analyze data to address research questions associated with each of the four evaluation areas. They conducted a user workshop, reviewed reports and internal TeraGrid documents, observed TeraGrid meetings, and participated on a TeraGrid Requirements Analysis team. Further, the research team conducted interviews with 86 individuals, including users of TeraGrid, developers of TeraGrid Science Gateways, TeraGrid personnel, non-TeraGrid users of HPC resources, and cyberinfrastructure experts. In addition, three surveys were conducted over the course of the study. These consisted of a survey of a sample of 595

current TeraGrid users and two surveys evaluating tutorials presented at the 2006 and 2007 TeraGrid conferences.

The TeraGrid evaluation research study was a separate activity from the planning process, but the results informed the steering committee's work. Publications that reported findings from portions of the investigation were reviewed by the committee (Krause & Zimmerman, 2007; Zimmerman, 2007; Zimmerman & Finholt, 2006; 2007). Other data collected as part of the evaluation study were in the process of being analyzed during the time period of the planning committee's work; preliminary findings were made available to the steering committee as appropriate.[5] Many of the results concurred with information collected as part of the planning process.

2.2.2 Planning Process Activities

A variety of methods were used to engage stakeholders, particularly current and potential users of TeraGrid, in the planning process and to gather information about their needs, ideas, and concerns. Information collection began in June 2007 and concluded in January 2008; Appendix B describes the activities in detail. Reports and summaries based on the information collected are available on the planning process web site.

2.2.3 Summary of Relevant Findings

In the course of its work, the committee learned much that was positive about the TeraGrid. We commend NSF for its vision in establishing it and recognize those involved in TeraGrid for their collaboration, hard work, and flexibility in response to organizational and technological changes. However, the findings from the evaluation research study and activities conducted as part of the planning process raised concerns that the committee felt should be addressed in the next version of the system. Many of these can be attributed to the experimental and evolutional nature of the TeraGrid and to the growing success of the facility. Such concerns included:

- Lack of clarity about what TeraGrid is and what it is not
  Users, especially potential users of TeraGrid, were not aware of the possibilities that TeraGrid could provide them and how they might go about gaining access to the resources. In addition, the dynamic nature of TeraGrid's growth and evolution and the need to balance multiple user needs hindered TeraGrid's ability to develop a shared project vision.
- Significant barriers to entry, including allocation process, time needed to train new users, and difficulty of use
  Even very experienced users with large allocations have ongoing needs for education and training as new graduate students and postdocs join their projects.

[5] The final report from the TeraGrid evaluation study was being completed at the same time as this document was prepared. Thomas Finholt and Ann Zimmerman of the UM-SI were PI and co-PI, respectively.

- Problems associated with relatively long queue times and heterogeneous hardware and software environments and different resource provider policies
  Long wait times in the queue were the main problem mentioned consistently across the data collection activities.
- Complex and changing software stacks, increasing the difficulty of porting and running codes
  In addition to the heterogeneity across sites in terms of policies, hardware, and software users noted the difficulty of keeping up with software changes.
- Single focus on very high-end computing instead of throughput computing
  While TeraGrid's focus should be on providing the most capable computing facilities, there are users, especially those who will gain access to TeraGrid through gateways, which require throughput capabilities and/or real-time access.
- Limited avenues for user involvement and input
  It was clear that users and other stakeholders desire more formal and informal avenues to provide input to TeraGrid.

We believe the NGRG can do even more to enable transformative research, and in the next section, we comment on six important and inter-related topics that underpin the basis of the findings for a path forward for TeraGrid that we present in section 4.0.

3.0 Key Issues

Our analyses of six areas, which we expand on below, form the basis of the committee's vision for an open, agile, production-quality NGRG focused on transformative research and an equal balance between technical and human capacity. These areas include:

- the science basis
- user base
- technical requirements
- program stability and robustness
- the role of the next generation research grid in the national and international cyberinfrastructure
- management and governance

3.1 Science Basis

Transformative research is not restricted by the scale or domain of research. The next generation research grid must strive to enable breakthrough research both in the sense of big science and little science. In addition, it must find ways to enable answers to questions researchers are currently unable to address as well as questions to appear in the next five years. While any individual project is not predictably transformative many fields have growing potential for such developments based on recent trends. These areas

depend on varying combinations of computational needs, data requirements, visualization, network services, and human capital, but all rely on aspects of each of these, which is why the NGRG must include and find a balance among them. In particular, we want to emphasize that human capital in the scientific community is as important as technical capability. Human capital includes many types of individuals such as users, developers, and educators as well as those who provide support, consultation, and training. Attention must also be paid to user community development, maintaining agility to respond to the needs of new fields that come online, mechanisms to balance infrastructure with research and development, and providing pathways through the cyberinfrastructure chain.

3.2 User Base

The user base is intimately related to the science basis. Discovery can come from many avenues, which is why the NGRG must be open and adaptable to new users, changing user requirements, and models of usage, including expansion from the current batch-oriented model to include interactive and on-demand processing. It also must provide user pathways to the high-end and the means to develop the human capacity to utilize the cyberinfrastructure. For example, users who only need to access data repositories connected to the NGRG (but no other NGRG resources) should be able to do so without obtaining an NGRG account. The "science gateway" model is a promising approach for providing wider access to complex resources and services for many research and education communities. The NGRG is focused on high-end uses and cannot directly support thousands of new users, but it should work closely with gateway developers to provide scalable approaches to broaden access. This is likely to require a separate effort on the part of NSF to determine how to best create, evolve, sustain, and evaluate gateway projects.

3.3. Technical Requirements

Users want a stable, production-quality environment. In the short-term, they want timely response; in the long-term they do not want constant change. In order to achieve this goal, tools and middleware need to be production hardened, easy to use, and ubiquitous, and the funding model must provide for long-term support of system and application software. The NGRG will broaden its scope to support a much larger and varied user community. To do so it must evolve its technical and management architecture so that it will support many more users across a wider range of scales, while retaining its role of providing access to the most capable cyberinfrastructure facilities. A potential approach for doing so is to explore the use of emerging services from the private sector, such as Amazon's Compute Cloud and IBM's Blue Cloud. Much more attention must be focused in areas related to data. Data storage, fusion, analysis, and visualization are critical to scientific discovery. The data infrastructure of the NGRG must be robust, persistent, ubiquitous, interoperable, and managed with consistent policies. Scalable tools for handling the enormous data sets produced by high-end systems are vital for maximizing the potential for knowledge creation. Finally, while the focus of the NGRG is on

production, there are likely to be needs for research and development platforms. These activities should take place in dedicated testbeds.

The science gateways, in particular, will be faced with computing throughput problems as their numbers of users increase. The NGRG must enable these science gateways to deliver throughput computing capabilities seamlessly. Virtualization capabilities and oversubscriptions of services will be a key element of success here.

3.4 Program Stability and Robustness

A robust and stable research infrastructure will promote sustained use and help to develop and maintain the human infrastructure that is critical to support and operate the next generation research infrastructure. The committee believes that it is scientifically viable and cost effective to fund cyberinfrastructure for longer periods than traditional research grants. Competition can promote innovation, but when used too frequently it is disruptive. There are precedents within NSF for longer term funding, and cyberinfrastructure warrants this same support. Additionally, there is no arbitrary number of resource providers that can be set for the next generation research grid. The number and characteristics of resource providers should be assessed initially as part of the planning grant activities and altered in periodic reassessments of the strategic plan (see Section 4). As part of this process, the committee believes that robust and stringent requirements should be set for the selection of resource providers. These requirements may vary according to the resource being provided (e.g., high end compute cycles, throughput compute cycles, data mining, interactive, non-scheduled visualization).

3.5 Role in National and International Cyberinfrastructure

While the NGRG is not *all* of cyberinfrastructure, it has multiple roles to play in national and international cyberinfrastructure. First, it will provide leading-edge cyberinfrastructure to the NSF constituency and proactively facilitate the use of its resources to carry out new as well as existing research projects. It will support the highest-end research through the most advanced and comprehensive computing, network, data storage, management, curation, and mining capabilities, and visualization, and discipline-specific services available. Further, it will integrate them with cyberinfrastructure from other agencies, nationally and internationally, as well as with regional and campus level cyberinfrastructures that support specialized and lower-end applications as well as education and training. Second, it should broaden its scope nationally so that there will be much greater participation both by users with many different needs and by providers of varied cyberinfrastructure resources. Third, it should assume a leadership role in creating a global grid research infrastructure by working with similar efforts that are underway in other countries. Here the emphasis must also be on production quality middleware infrastructure, rather than research oriented prototypes. "The Grid" as originally envisioned does not exist today, unless a user is very knowledgeable and willing to surmount significant barriers to compute in this fashion. Lastly, the NGRG should work with the relevant scientific communities to explore

establishing agreements with other agencies about sharing data and other cyberinfrastructure resources, as well as developing middleware.

3.6 Governance and Management

Effective and transparent governance of the NGRG will be essential. The NGRG will be a national scientific resource to provide services and infrastructure to the scientific community and contribute to scientific progress across fields and institutions. The governance and management of a distributed system that crosses institutional and administrative boundaries is complex and challenging, but it is also critical to the success of the NGRG. The Institute on Governance defined governance as "a set of ideas about how direction is provided to human activity." [6] Management consists of executive decision-making and implementation within the framework established by governance. The main feature that distinguishes governance from management is that the former is concerned with how the *big* (or strategic) decisions are made and who makes those decisions.

There are many different governance models, but in general, transparency and the sustained involvement of key stakeholders in an independent advisory board are necessary conditions for effective governance. Further, governance and management of research infrastructure is appropriately different from research grants. In the past, NSF has pioneered the creation of boards of overseers and advisory structures for major research infrastructure projects. The NGRG will need to draw on this tradition to obtain the necessary strong leadership and management on behalf of the larger research community. This will require the creation of formal channels for the NSF and the NGRG management to obtain ongoing input from the user and other stakeholder communities.

4.0 Findings

The TeraGrid has been opportunistic in responding to increases in the number of resource providers, the evolving technological landscape, and changes in the types of users, usage modes, and user requirements. However, the open, agile, and robust production infrastructure that we envision for the next generation research grid requires:

- a funding model designed to support the program attributes over extended time periods
- a strategic plan that includes statements of vision, mission, and values, a list of specific goals, a description of the ways in which those goals will be met, and scheduled reassessment of the plan at set intervals regularly against pre-specified metrics, and
- a governance structure and a management plan that includes multiple avenues for stakeholder participation, including a formal advisory structure that reports both to the NSF and to project management.

---

[6] For further information, see the web site of the Institute on Governance: http://www.iog.ca/

To ensure that the NGRG has the stability, direction, leadership, and community support that will be necessary to its success and to its ability to remain agile in the face of technological change, we suggest that the NSF prepare a two-step announcement of opportunity for competitive planning grants leading ultimately to the selection of **an entity to manage the NGRG**. The initial announcement would require proposers to describe how they would conduct a process whose end result would be:

- a strategic plan, including a description of a strong and responsive governance structure and management plan (as outlined above)
- a description of how standards would be used in the creation of the next generation research grid and the way in which the management structure would effect their use
- plans to create an accessible and user-friendly production quality cyberinfrastructure environment
- plans to provide mechanisms and procedures to enable research, development, and testing of cyberinfrastructure standards and tools, without impacting production operations of the NGRG
- strategies for broadening participation to include new users, disciplines, resource providers, partners, and science gateways
- processes to interoperate with Track 1 and Track 2 systems and with data storage, analysis, and visualization systems and pathways to these resources from campus level systems and other high-performance computing centers in an extensible partnership mode[7]
- an approach to preserve agility in the face of inevitable technological change
- plans for the career development of people who will support the infrastructure and computational science
- a plan to coordinate and cooperate with other national and international cyberinfrastructure providers and to provide leadership in the development of an international grid infrastructure
- an allocation process matched to program attributes
- an education and outreach program serving and expanding the community that will create, utilize, support, and extend the cyberinfrastructure to enable research discovery and learning for present and future generations

[7] The Track 1 system is a petascale computer. The University of Illinois at Urbana-Champaign will receive funding to acquire and make available a petascale computer it calls "Blue Waters." The system is expected to go online in 2011. NSF's Track 2 initiative is a four-year activity designed to fund the deployment and operation of up to four leading-edge computing systems. These systems will be integrated into TeraGrid (NSF, 2007).

The announcement might require a letter of intent, followed by a pre-proposal process and an invitation to multiple proposal teams to submit full planning proposals. The pre-proposals would be reviewed by a panel of experts and the submitters of the most promising pre-proposals would be awarded 6-8 month planning grants that would enable them to develop a document that includes the components described above and that is based on widespread participation of the research, high-performance computing, software and middleware, networking, data, education and training, and other stakeholder communities. We envision the grantees holding workshops and conducting other activities to engage the stakeholder communities. At the end of the planning period, the grantees would submit full proposals for developing and managing the NGRG. These proposals would be reviewed by a panel that includes experts in infrastructure as well as researchers from multiple user domains and others with expertise critical to the program's success such as the management of large community facilities. The agreement with the selected team would last for a minimum of five years and would include regularly scheduled assessments against pre-specified metrics and options for renewal of the agreement.

On the NSF side, it is vitally important that OCI and the directorates coordinate their strategic plans because in many ways the directorates are the most important customers of the next generation research grid. The NGRG should be strongly driven by the needs of current and future research communities able to make significant strides with the use of modern high-end cyberinfrastructure. Thus, in addition to the product that would result from the announcement of opportunity, mechanisms should be developed for other research programs supported by NSF directorates to coordinate with OCI and to make use of the NGRG as an integral part of their programs and to provide incentives for alignment.

Summary

We strongly endorse a next phase for the TeraGrid program, which we have referred to in this report as the *Next Generation Research Grid*. The open, agile, and robust production infrastructure that we envision for the NGRG requires a funding model to support program attributes over extended time periods; a strategic plan and scheduled reassessment of the plan at set intervals; and a transparent governance structure and management plan that includes multiple avenues for stakeholder participation. To ensure that the NGRG has the stability, direction, leadership, and community support that will be necessary to its success and to its ability to remain agile in the face of technological change we suggest that the NSF prepare a two-step announcement of opportunity for competitive planning grants leading ultimately to the selection of **an entity to manage the NGRG**. We hope that an additional outcome will be a new name that reflects the vision of this open, high-end, production quality infrastructure. The agreement with the selected team would last for a minimum of five years and would include regularly scheduled assessments against pre-specified metrics and options for renewal of the agreement. In addition, to the product that would result from the announcement of opportunity, it is important that mechanisms be developed for other research programs

supported by NSF directorates to coordinate with OCI and to make use of the NGRG as an integral part of their programs and to provide incentives for alignment.

Acknowledgments

We are pleased to acknowledge the many people who contributed to our work, although we accept all responsibility for the contents of this report. We thank the many individuals who participated in planning events, including the summer 2007 planning workshops and the town hall meetings, and who contributed position papers or provided comments and ideas in other ways. We appreciate their combined time, energy, and thoughtfulness. The SC '07 Birds of a Feather session would not have been possible without the assistance of Janet Brown, Pittsburgh Supercomputing Center, Nancy Wilkins-Diehr, San Diego Supercomputing Center, and Harvey Wasserman, Lawrence Berkeley National Laboratory. Finally, we wish to acknowledge the invaluable support provided by Veda Emmett, Executive Assistant to Tim Killeen, and Rebecca O'Brien and Cheng-Lun Li, University of Michigan School of Information.

References

Krause, M. and Zimmerman, A. (2007). *TeraGrid '06 Tutorial Evaluation*. Collaboratory for Research on Electronic Work, School of Information, University of Michigan, Ann Arbor, MI.

Langer, J. S., ed. (1998). *National Workshop on Advanced Scientific Computation*. National Academy of Sciences.

Lawrence, K. A. and Zimmerman, A. (2007a). *TeraGrid Planning Process Report: August 2007 User Workshops*. Collaboratory for Research on Electronic Work, School of Information, University of Michigan, Ann Arbor, MI. Retrieved May 5, 2008, from http://www.teragridfuture.org/system/files/TeraGrid+User+Workshops+Final+Report.pdf

Lawrence, K. A. and Zimmerman, A. (2007b). *TeraGrid Planning Process Report: June 2007 Workshop for Science Gateways*. Collaboratory for Research on Electronic Work, School of Information, University of Michigan, Ann Arbor, MI. Retrieved May 5, 2008, from http://www.teragridfuture.org/system/files/TeraGrid+Science+Gateways+Workshop+Report.pdf

National Science Foundation. (2007, August 8). *National Science Board approves funds for petascale computing systems*. Press Release 07-095. Retrieved May 5, 2008, from http://nsf.gov/news/news_summ.jsp?cntn_id=109850&org=NSF&from=news

National Science Foundation, Cyberinfrastructure Council. (2007). *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, Arlington, VA.

National Science Foundation, Directorate for Computer and Information Science and Engineering, Division of Advanced Computational Infrastructure and Research. (2001). *Distributed Terascale Facility (DTF), Program Solicitation (NSF 01-51).* National Science Foundation. Retrieved May 5, 2008, from http://www.nsf.gov/pubs/2001/nsf0151/nsf0151.htm

President's Information Technology Advisory Committee. (1999). *Information Technology Research: Investing in Our Future*. National Coordination Office for Computing, Information and Communications, Arlington, VA.

Reed, D. A., Patrick, M. L., Sugar, R., Keyes, D., & Voigt, R. (1998). *Terascale and Petascale Computing: Digital Reality in the New Millenium.*

Zimmerman, A. (2007). A socio-technical framework for cyberinfrastructure design. *e-Social Science Conference, Ann Arbor, MI, October 7-9, 2007.*

Zimmerman, A. and Finholt, T.A. (2007). Growing an infrastructure: The role of gateway organizations in cultivating new communities of users. Pages 239-248 in *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (GROUP '07)*, ACM Press, New York, NY.

Zimmerman, A. and Finholt, T. A. (2006). *TeraGrid User Workshop Final Report.* Collaboratory for Research on Electronic Work, School of Information, University of Michigan, Ann Arbor, MI. Retrieved May 5, 2008, from http://www.crew.umich.edu/research/teragrid_user_workshop.pdf

Appendix A: Planning Process Leadership and Facilitation

TeraGrid Planning Process Steering Committee

- Timothy Killeen (Chair), National Center for Atmospheric Research
- Roberta Balstad (Associate Chair), Columbia University
- Arthur (Buddy) Bland, Oak Ridge National Laboratory, National Center for Computational Sciences
- Roscoe Giles, Boston University
- Myron Gutmann, Inter-university Consortium for Political and Social Research, University of Michigan
- Gwendolyn Huntoon, Pittsburgh Supercomputing Center
- Gerhard Klimeck, Purdue University
- Paul Messina, Caltech and Argonne National Laboratory (retired)
- B. Montgomery Pettitt, University of Houston
- Edward Seidel, Louisiana State University
- Joan-Emma Shea, University of California, Santa Barbara
- Alexander Szalay, The Johns Hopkins University
- Dane Skow, TeraGrid Ex-Officio Member

TeraGrid Planning Process Facilitation Team

- Ann Zimmerman, PI of the Facilitation Grant for the TeraGrid Planning Process, University of Michigan School of Information
- Katherine Lawrence, University of Michigan School of Information
- Peter Backlund, National Center for Atmospheric Research

Appendix B: Information Gathering and Dissemination

The steering committee's findings are based on a series of activities that were designed to engage current and potential users of TeraGrid and to provide information on the planning process. We considered the results of three invitational workshops conducted by UM-SI, hosted "town hall" meetings at professional conferences, solicited position papers from current TeraGrid users and other national and international stakeholders, examined relevant reports and documents, and used the planning process web site to disseminate and collect information. Besides these formal activities, we and other committee members had conversations and written exchanges with interested stakeholders that further informed our findings.

Workshops

The first workshop invited participation from all of the TeraGrid Science Gateways, and seventeen of the twenty-one gateways attended. The second and third workshops included current and potential users from research disciplines that use high-performance computing and grid resources (Lawrence & Zimmerman, 2007a; 2007b). The workshops were conducted by UM-SI in advance of the committee's formation.

Town Hall Meetings

The "town hall" meetings, hosted at the SC07 (supercomputing) conference and the 2007 American Geophysical Union (AGU) Fall Meeting, provided a venue for interested individuals to share their comments, questions, and concerns about the future of TeraGrid and the planning process.

Position Papers

Our call for position papers generated almost twenty formal responses from individuals and groups in the United States and elsewhere around the world and a similar number of informal input via email.

Existing Documentation

We reviewed documents related to the background, history, and operations of TeraGrid. We also read reports and reviewed web sites on national and international grids and other relevant topics.

TeraGrid Future Website

The UM-SI created a website (www.teragridfuture.org) to provide a dissemination point for information and reports generated by the planning process as well as to collect comments from website visitors.

TeraGrid Future Newsletter

We produced an electronic newsletter with updates on the planning process and information on events.

Appendix C: Letter to Stakeholders Regarding Draft Report

June 5, 2008

Letter to Stakeholders Regarding the Draft Report of the TeraGrid Planning Process Steering Committee

Background

This letter concerns the draft report from a project funded under NSF award OCI-0724300 to work with stakeholder communities to collect ideas for the next generation of the TeraGrid. The National Science Foundation awarded the grant to the University of Michigan's School of Information to facilitate a planning process to help guide the future evolution of TeraGrid. As the Principal Investigator of the planning grant, I convened a steering committee, representative of key stakeholder communities and with diverse expertise to help achieve the project goals.

The draft report of the steering committee reflects the consensus view of its members and is based on extensive information gathering and deliberation, most of which occurred between mid-September 2007 and the end of January 2008. I am soliciting your comments on the draft report.

This is a draft issued to receive input; IT IS NOT A FINAL DOCUMENT. As such, it is important that the header on the document be respected and that the draft report is not quoted or considered final. An expected release date of the final report will be established once the steering committee completes a review of the comments and considers revisions.

Submittal of Comments

The steering committee welcomes your constructive input on any aspect of the draft report. To ensure that your comments are properly considered, it is important that they be submitted in a format that will facilitate review by the steering committee. Therefore, they should be specific to the maximum extent possible; they should clearly state the topic and location (page number, paragraph or sentence). The template that appears at the end of the draft report can be used to help you organize your comments. This form, and other information regarding the TeraGrid planning process, is also available on the planning web site: http://www.teragridfuture.org.

All comments must be received by 5 p.m. EDT on Tuesday, July 8, 2008; those received after this time will not be considered. Attributed comments are preferred, but anonymous comments will also be accepted. If you do not wish your comments to be made public, please note that when you submit them. Comments may be sent via e-mail to **sc-report-comments@umich.edu** or faxed or mailed to Dr. Ann Zimmerman, UM School of Information, 1075 Beal Ave., Ann Arbor, MI 48109-2112; fax 734-764-1555. A public summary of the comments received will be prepared when the final report is issued.

On behalf of myself and the steering committee, thank you for your input into this important process.

Sincerely,

Ann Zimmerman

Ann Zimmerman, PhD
PI, TeraGrid Planning Process
Research Assistant Professor
UM School of Information

Appendix D: Template for Comments on Draft Report

The steering committee welcomes your constructive input on any aspect of the draft report. To ensure that your comments are properly considered, it is important that they be submitted in a format that will facilitate review by the steering committee. The template below is provided to assist you in the submittal of your comments.

Comments are due **by 5 pm EST on July 8, 2008** and may be submitted to the following:

E-mail: sc-report-comments@umich.edu

Fax: Attn: Dr. Ann Zimmerman, 734-764-1555

Mail: Dr. Ann Zimmerman, University of Michigan, School of Information, 1075 Beal Ave., Ann Arbor, MI 48109-2112 USA

****

Name (optional) _____

Affiliation (optional) _____

Contact: Phone number or e-mail address where you can be contacted in case of questions. (Optional) _____

☐ Please check this box if you do not wish your comments to be attributed to you publicly.

Report section	Page number	Paragraph/ sentence	Comment

Appendix E: Summary of Comments Received on Draft Report

Two individuals submitted comments on the draft report.

Dick Repasky (Indiana University) submitted feedback using the template for comments (see Appendix D). His comments included editorial suggestions and remarks related to the subject matter presented in the report; the latter are summarized below.

- The barriers to using TeraGrid are not clearly stated or adequately described. For example, the first paragraph in the first section of the report should be revised to state that TeraGrid is impossible to use without investing large amounts of time to learn how to apply for an allocation, log in, and compile and run code.
- A grid requires a metascheulder and site-independent methods for storing and retrieving data The word "grid" should be dropped from the name for the next phase of TeraGrid (i.e., NGRG) unless these requirements are met.
- To create cyberinfrastructure is to create an arrangement of computing resources. The issues that can be addressed are those that are affected by the choice of and arrangement of computing resources. For example, if the scientific problem requires a bigger parallel computer, it may be possible to tie together distant computers in a manner in which they behave like a bigger computer. Or, it may be possible to improve the speed with which data are delivered to a running program.
- Outside of the realm of what can be solved by the design of cyberinfrastructure are many of the issues that commonly thwart the progress of science: simplying assumptions that render problems untractable, new algorithms, and new conceptual frameworks for casting problems.
- The tension between cyberinfrastructure and research and development deserves more space and a firm recommendation.
- It is important to ask whether broadening the scope to include many users with many different needs witll reduce the "average high-endedness" of cyberinfrastructure.

Comments from Vijay Agarwala (Pennsylvania State University), which are summarized below, were based on a white paper he wrote and on an article that appeared in *HPCwire*.

- While Track I and Track II grants will lead to major advances in science in the years to come, there is also a need for up to twenty "Track III" centers awarded competitively to U.S. universities.
- The rationale for creating these new Track III centers is: 1) to increase global competitiveness of U.S. industry across all sectors, and 2) to reduce or reverse the outsourcing of engineering services that has been taking place at an alarming pace.

Further details regarding Dr. Agarwala's recommendations and rationale are available in the January 2008 *HPCwire* article: http://www.hpcwire.com/features/17908929.html.