Report from the TeraGrid Evaluation Study, Part 1: Project Findings

Ann Zimmerman

Thomas A. Finholt

August 2008

Collaboratory for Research on Electronic Work
School of Information
University of Michigan
1075 Beal Avenue
Ann Arbor, MI 48109-2112
http://www.crew.umich.edu

Table of Contents

Executive Summary

TeraGrid integrates multiple high-performance computing resources at distributed provider facilities. In 2006, the National Science Foundation (NSF) awarded a grant to the University of Michigan's School of Information (UM-SI) to conduct an external evaluation of TeraGrid. The primary goals of the evaluation were to provide specific information to TeraGrid managers that will increase the likelihood of TeraGrid success, and to give NSF and policy makers general data that will 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
- satisfaction among TeraGrid partners

We employed a mixed method approach that consisted of a user workshop; participant observation; document analysis; interviews with 86 individuals representing five different categories; a survey of a sample of 595 TeraGrid users; and two surveys to assess TeraGrid tutorials held in 2006 and 2007. Most of the data were collected from June 2006 through May 2007.

Findings from the evaluation study are presented in two parts. In this first part, we report results from analyses of all data collected during the investigation. Detailed findings from the user survey are presented in Part 2 of the report.

TeraGrid as a Virtual Organization

A virtual organization (VO) is distributed across space and time, dynamic in processes and structure, and enabled and enhanced by technologies for communication and coordination. As a VO, TeraGrid has the potential to capitalize on the advantages of organizing in this way, but it also confronts many of the challenges faced by other distributed organizations. The TeraGrid VO can be characterized as a series of tensions that both create challenges and help the organization to grow, innovate, and adapt.

One set of tensions arises from the diverse characteristics, histories, and cultures of the resource provider sites that participate in TeraGrid, their need to collaborate as well as compete, and their desire to retain individual autonomy and identity while cooperating under the TeraGrid banner. In addition, as new institutions join the TeraGrid scaling becomes a challenge. For instance, communication and coordination, management and governance, and technical integration become more complex. A second source of stress stems from TeraGrid's three-part mission to support the most advanced computational science in various domains, to empower new communities of users, and to provide resources and services that can be extended to the broader cyberinfrastructure. Each of these goals can require different methods, resources, expertise, and strategies to achieve, making it difficult to establish and balance priorities and to define success. Further, some goals fit better with the existing

strengths, customer bases, and priorities of individual resource providers than others. Finally, there is tension between the desire for infrastructural reliability and stability, particularly on the part of users, and the research and development that are often necessary to create a distributed cyberinfrastructure. TeraGrid has developed approaches to deal with some of these challenges, which can provide valuable lessons to other VOs.

Prior to TeraGrid, there was limited opportunity for interaction between personnel at the different resource provider facilities. Project participants noted that collaboration with people at other sites has led to the development of new technologies, processes, and systems to support distributed computing; made them more aware of the capabilities and expertise at other sites, which has helped to improve service to users and enhance local procedures; and opened up TeraGrid resources and services to new user communities. While they viewed these as important successes, many believed that a clearer project vision would direct efforts in ways that would better serve users' needs.

Grid Computing

The original notion of grid computing as running at more than one site concurrently (i.e. co-scheduling) has evolved into the idea of employing *the grid* to enable a variety of distributed tasks and modes of usage. This is viewed by most interviewees as a positive development.

User Needs Analysis

The goal of the user needs analysis was to understand the needs of different types of users and the factors that influence user behavior. The results from the user survey offer a generalizable view of current TeraGrid users while the qualitative data provide in-depth information and help explain some of the survey findings.

TeraGrid users are numerous, diverse, and distributed. Finding meaningful ways to classify users can improve strategies for delivering TeraGrid support and consultation services to users; inform decisions related to hardware, software, and policy; and provide mechanisms for users to interact with each other in support of their needs. Our results shows that experience level; algorithms, models, and software; usage mode; and the relationship between the nature of the research problem and the technical infrastructure are potentially productive means to distinguish the needs of different users. In addition, we found that usability issues affect users of all types and experience levels and that problems with lengthy queue times go beyond user irritation and inconvenience and affect the efficiency and speed at which science is conducted.

The resources available through TeraGrid are critical to most users. Although batch use of a single resource at one or more sites is a common mode of usage, users noted that the ability to move data from one computer to another was an important aspect of TeraGrid. They also noted that the use of TeraGrid resources made it possible for them to simulate phenomenon at longer time scales or across a continuum and to improve experimental and engineering designs. Future requirements will be shaped by users' needs to manage, store, and analyze growing amounts of data and to make existing codes run efficiently on more capable

resources. Education and training, tools to support collaboration, and access to a wider variety of usage modes will also be increasingly important.

TeraGrid Science Gateways

Recognizing that many disciplinary communities were building elements of their own cyberinfrastructure, TeraGrid formed partnerships with other projects to provide TeraGrid resources and services to user communities through tools and services they were already using. One goal of gateways is to enable entire communities of users associated with a shared research goal to use TeraGrid resources through a common interface. Developers of science gateways have two roles: 1) TeraGrid user, and 2) intermediary between user communities and TeraGrid.

As users of TeraGrid, developers of science gateways are in need of tools and processes that help make development easier and that assist them to support their users. These include basic services that gateways can use instead of creating or hosting their own; templates and standardized systems; and standardization to support the effective use of allocations and meta-scheduling. As intermediaries between the needs of TeraGrid and their users, science gateways appear to play an important role in introducing TeraGrid to potential users and enabling current users to conduct their work in new ways.

1. Introduction

TeraGrid integrates multiple high-performance computing resources at distributed provider facilities. In 2006, the National Science Foundation (NSF) awarded a grant to the University of Michigan's School of Information (UM-SI) to conduct an external evaluation of TeraGrid. The primary goals of the evaluation were a) to provide specific information to TeraGrid managers that will increase the likelihood of TeraGrid success, and b) to give NSF and policy makers general data that will 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
- satisfaction among TeraGrid partners

Findings from the TeraGrid evaluation study are presented in two parts. In this first part, we report results from analyses of all data collected during the investigation. As appropriate, findings from the *TeraGrid User Survey* we conducted in late 2006 are integrated into this document; detailed findings from the survey are presented in Part 2 of the report.

The primary audience for this report is NSF personnel, particularly those responsible for management of NSF's cyberinfrastructure investments and those in divisions that support the research of communities that are current or future users of TeraGrid. The other main audience for this report is TeraGrid itself, specifically senior managers, but also all individuals at the TeraGrid resource provider (RP) sites who are or have been involved in the project whether formally or informally. The document was also written with many other TeraGrid stakeholders in mind such as TeraGrid users, including individuals affiliated with projects that provide gateways to TeraGrid resources and services, developers of software and applications, and other providers of high-performance computing (HPC) or grid services in the United States and internationally. Given the short time period of the study, our evaluation was necessarily formative, and our primary goal was to provide feedback to NSF and to TeraGrid that would help with future planning. We hope the findings will be a significant step toward filling gaps in our knowledge about the needs of scientists and engineers who use NSF-supported HPC resources and services, particularly TeraGrid, the factors that affect their use, and the impact of this use on research outcomes. In addition, we aim to improve understanding of the organizational, social, and technical challenges associated with the delivery of cyberinfrastructure (CI) via a virtual organization such as TeraGrid.

We begin this report with a description of the evaluation research questions and the multiple methods used to collect and analyze the data on which the study's results are based. Next, we provide an overview of the history, current organization, and operations of TeraGrid. In addition to providing background for those who are less familiar with TeraGrid, this

information is relevant to findings discussed later in this document. The majority of the report focuses on the study's findings. The results are divided into sections that cover the main topics of the investigation. The end of each results section includes a summary and a brief discussion of potential implications. Given the formative nature of the evaluation, the latter should be viewed as preliminary and are provided for the purpose of offering some discussion of the results. We conclude the report with a discussion of the study's limitations and needs for future research.

2. Research Questions

As stated in the previous section, the evaluation study was designed to address questions related to user needs, experiences, and satisfaction; the impact of TeraGrid on research outcomes; quality of TeraGrid education, outreach, and training activities (EOT); and collaboration between TeraGrid resource providers.

Questions related to user needs and requirements:

- What factors affect users' computing needs and requirements?
- How are the needs of users expected to change over the next five years?
- What factors affect users' behavior as it relates to their use or non-use of computing resources and services?

Questions related to impact:

- How does the use of TeraGrid impact the work of scientists and engineers?
- What are the benchmark factors for impact in particular areas of research?

Questions related to TeraGrid as a virtual organization:

- What factors affect satisfaction among the resource providers?
- How is information shared across sites?
- What mechanisms are used to coordinate activities and services?

Question related to EOT activities:

- How do attendees of EOT events perceive the quality of instruction and instructional materials?
- How likely are attendees to participate in future tutorials or workshops?
- What is the level of enthusiasm for TeraGrid use before and after participation in an EOT event?

We investigated the questions above using multiple research methods. We describe these approaches in the next section.

3. Research Methods

Mixed methods research involves the collection, analysis, and integration of quantitative and qualitative data in a single study. This form of research can provide a better understanding of a question or topic than is possible with a single research approach

(Creswell, 2003). In this particular study, the collection, integration, and analysis of both qualitative and quantitative data improved validity, credibility, and comprehensiveness of the findings.

3.1 Data Collection

Most of the data collection activities occurred from June 2006 through May 2007. Institutional Review Board (IRB) approval was obtained from the University of Michigan for the study. In keeping with the conditions of IRB approval and to protect the anonymity of research subjects, we provide only general information regarding the sites we visited and the individuals we interviewed.

3.1.1 User Workshop

The first major activity we conducted as part of the evaluation study was a June 2006 user workshop whose purpose was to begin to examine the relationship between TeraGrid's development priorities and the needs of its users. The invitation-only workshop was funded by TeraGrid and held in conjunction with the first TeraGrid Conference in Indianapolis, Indiana. The workshop was designed to assess user requirements from the standpoint of how TeraGrid could enable the next big breakthrough in workshop attendees' lines of research. Twelve individuals, representing the fields of biology, chemistry, geosciences, physics, and social and behavioral sciences, participated in the event.

Two key conclusions were generated from the workshop. First, we found that there are a range of technical needs even in a group of users who have mostly large allocations. Second, although the technical requirements of users vary, the social and educational challenges they face are similar. These two findings provided a preview of results that emerged from the study overall. Detailed results from the user workshop are available in a separate report (Zimmerman & Finholt, 2006).

3.1.2 Site Visits

From mid-July through early October 2006, we visited seven sites around the United States with the primary purpose of conducting interviews with a variety of individuals; further detail on interviewees is provided below. Five of the cities we visited were home to a TeraGrid resource provider site as well as to users or other interviewees. In addition to the opportunity to conduct face-to-face interviews with a large number of people in a short amount of time, the site visits also allowed us to observe users in their working environments and to learn more about the operations of TeraGrid RP sites.

3.1.3 Interviews

We conducted interviews with 86 individuals in 66 separate interview sessions. As these numbers indicate, most interviews were conducted with a single person, but several took place with groups of two or more individuals. The individuals we interviewed were affiliated with 13 different institutions, including five RP sites, as noted above. At least

one institution from each major geographic area of the United States was represented in the sample.

The interview sessions totaled more than 47 hours with individual interviews averaging 43 minutes in length. Most interviewees agreed to allow the interview to be audiotaped, and transcriptions were made from the digital audiofile. When an individual declined to have the interview recorded the interviewer took notes instead.

We conducted interviews with individuals who represent five different categories that we identified in advance as being important to address the objectives of the evaluation study. Table 1 summarizes the number and types of people interviewed, and the text below describes each category in further detail.

Category	Definition	Number of interviewees
TeraGrid Users • Individual Researchers	Individuals associated with a project that had a TeraGrid allocation at the time of the interview	26
TeraGrid Users • Science Gateway Developers	Individuals who on a day-to-day basis spend some portion of their time working on a project designated as a TeraGrid Science Gateway	27
TeraGrid Personnel	Individuals employed by one of the TeraGrid RP sites who have a formal or informal role in the TeraGrid project	26
Non-TeraGrid Users of HPC Resources	Individuals who use HPC computing resources other than TeraGrid	3
Cyberinfrastructure Experts	Individuals with extensive knowledge of high-performance computing	4

Table 1: Number and Types of People Interviewed

Several interviewees belonged to more than group. For example, an individual working on one of the science gateways was a domain scientist, who had a TeraGrid account to support his own research. In these cases, we asked respondents questions related to the main reason for which they we selected for an interview, although comments they made relevant to their other roles and experiences were included in our analyses.

TeraGrid Users

We interviewed two types of TeraGrid users: 1) individual users, and 2) science gateway developers. Individual users (n=26) included principal investigators (PIs), faculty and research scientists, postdoctoral associates, and graduate students with a TeraGrid account. These users were affiliated with seventeen different projects and eleven institutions. Table 2 provides information on the research area, position, and the largest allocation the interviewee's project had at the time. When proposals are accepted and projects are granted an allocation, they are assigned to one of three award categories: development allocations (DACs), medium resource allocations (MRACs), and large resource allocations (LRACs). The awards differ based on the number of service units allotted, ranging from 30,000 for DACs, between 30,000 and 200,000 service units for MRACs, and over 200,000 service units for LRACs. Although LRAC awardees are small in number relative to all awards made, they represent a major portion of the use of TeraGrid allocated resources (see section 8.1.1). Whereas DAC awardees are many, they use a small percent of allocated resources. In the interviews, we focused on LRAC and MRAC users because they represent the majority of TeraGrid use. In addition, since some LRAC and MRAC users sometimes also have a DAC award, it was challenging to find users who had only a DAC award and were available for an interview. The purpose in interviewing DAC only users was to learn about the needs of individuals who were new to TeraGrid. We were able to gain insight into these needs in other ways, particularly by talking with graduate students and other researchers associated with LRAC and MRAC projects, who had not used TeraGrid before, and through the *TeraGrid User Survey* (see Part 2 of the report).

We asked individual users questions about their research, including future research goals, their experiences using HPC resources generally and TeraGrid in particular, including when they started using them, how they learned to use them, how they obtained help, what resources and capabilities they used (i.e., mass storage, visualization, computation), and what resource provider(s) they used, including facilities other than TeraGrid. We also asked interviewees to describe the impacts of TeraGrid/HPC on their research. Finally, we sought information about the other technologies they used, how they spent their time (i.e., administration, research, teaching), and who they collaborated with. Our interview protocol assumed that users understood what TeraGrid *is* in the sense of both its organization and purpose. We soon found, however, that users' knowledge and perceptions about TeraGrid varied widely, so we added a question to the protocol that asked interviewees to define TeraGrid from their point of view. If users had a vague notion of TeraGrid, we framed our subsequent questions around their use of NSF HPC resources more generally rather than TeraGrid specifically.

Project	Position	Research Area	Allocation Level
A	Principal Investigator	High-Energy Physics	LRAC
B	Doctoral Student	Chemistry	LRAC
C	Principal Investigator	Biochemistry	LRAC
D	Principal Investigator	Materials Science	LRAC
D	Principal Investigator	Materials Science	LRAC
E	Research Scientist	Molecular Biology	LRAC
F	Research Programmer	Molecular Biology	LRAC
G	Postdoc 1	Chemistry	LRAC
G	Postdoc 2	Chemistry	LRAC
H	Principal Investigator	Earth Science	LRAC
I	Doctoral Student	Chemistry	LRAC
J	Research Scientist	High-Energy Physics	LRAC
K	Masters Student	Chemical Engineering	LRAC
L	Principal Investigator	Astrophysics	MRAC
L	Doctoral student	Astrophysics	MRAC
M	Principal Investigator	Economics	MRAC
M	Postdoc	Economics	MRAC
M	Doctoral Student	Economics	MRAC
N	Principal Investigator	Geoscience	MRAC
N	Postdoc	Geoscience	MRAC
N	Doctoral Student	Geoscience	MRAC
O	Principal Investigator	Molecular Biology	MRAC
P	Principal Investigator	Chemistry	MRAC
P	Doctoral Student	Computer Science	MRAC
Q	Research Scientist	Geoscience	DAC
Q	Research Programmer	Computer Science	DAC

Table 2: Individual User Interviewees

Developers of science gateways are a second type of TeraGrid user. Here, we use the term *developer* broadly to mean people who on a day-to-day basis spend some portion of their time working on a project designated as a TeraGrid Science Gateway.[1] Early in its history,

[1] A project is defined by TeraGrid as a TeraGrid Science Gateway if it has an allocation on the TeraGrid.

TeraGrid conceived the idea for what has become the TeraGrid Science Gateway program. Recognizing that many disciplinary communities were building elements of their own cyberinfrastructure, TeraGrid set out to form partnerships that would provide TeraGrid resources and services to user communities through tools and services they were already using (Catlett, Beckman, Skow, & Foster, 2006). TeraGrid's role is as a back-end service provider with the gateway serving as the front end to the user. There were approximately 20 such projects during the time of the evaluation, and we interviewed individuals affiliated with eight of them. Interviewees (n=27) included persons with expertise in the scientific domain represented by the project, technologists developing and/or making applications and services available through the gateway, project management experts, and education and outreach professionals. We provide few details regarding the gateways we studied since this information could make it possible to identify the projects and thus compromise the anonymity of our respondents. We can state that we investigated gateways that serve a diverse set of disciplines. The questions we asked interviewees varied depending on their role in the project, but in all cases we spoke with at least one person who was able to tell us about the motivation for the gateway, especially the scientific needs that the project was trying to fulfill, the gateway's intended user community, the kinds of resources and capabilities that were perceived as being important for the user community, and the impacts that the gateway would have if it were successful. We also asked more specifically about the need for HPC resources and services within the gateway and experiences in working with TeraGrid.

TeraGrid and Centers Personnel

The second main category of interviewees is individuals who were employed by one of the TeraGrid RP sites and who had a formal or informal role in the project. We interviewed 26 people at five RP sites whose salary was supported in whole or in part by TeraGrid funds or, who received no funding from TeraGrid, but who had direct involvement in the project such as supervising people supported by TeraGrid funds. We refer to both types of people as TeraGrid personnel. The individuals we interviewed worked in or were responsible for a variety of areas such as operations, networking, security, user support, systems administration, education, external affairs, and software development. In addition, we interviewed both staff and managers. The latter included a sample of individuals who serve as Area Directors or are members of the Grid Infrastructure Group (GIG) or the Executive Steering Committee (ESC) (see section 4.2). Most interviewees' experiences with TeraGrid dated back to the start of their institution's affiliation as a member of TeraGrid. The questions we asked TeraGrid personnel varied slightly depending on their role in the project, but in all cases we asked individuals about their responsibilities, how and when they came to be involved in the project, their opinions regarding TeraGrid's challenges and successes, and other TeraGrid personnel they work with and the ways in which those interactions occur.

Non-TeraGrid Users of HPC Resources

We interviewed three individuals who use HPC facilities other than TeraGrid to understand why they do not use TeraGrid. Due to the small sample size, few conclusions can be drawn from these interviews, but they were helpful in gaining additional insight into the factors that influence researchers' computing decisions. The questions we asked non-TeraGrid users of HPC resources were similar to those we asked individual TeraGrid users.

Cyberinfrastructure Experts

Finally, we interviewed four people that we describe as cyberinfrastructure experts. These individuals have been involved in the world of high-performance computing in various capacities for many years. Among the interviewees in this category were a former funding agency program manager, the director of a campus computing center, a former employee of Internet2, and a computer science researcher employed by a university computing center.

3.1.4 Participant Observation

Participant observation enables a researcher to gain close familiarity with a group or community and their practices through intense involvement with them, often over a long period of time. In this study, participant observations included our attendance at five TeraGrid quarterly management meetings during the period September 2005 through April 2007. The primary purpose of the face-to-face TeraGrid quarterly meeting is to bring together the RP-PIs, the ESC, and the GIG (including Area Directors) to review and assess progress over the past quarter and to make necessary adjustments in plans, priorities, or projects. Representatives from some science gateway projects also participate in these meetings. The September 2005 meeting was the first face-to-face gathering of the TeraGrid principals following NSF's $150 million, 5-year extension of the TeraGrid for the period August 1, 2005-July 31, 2010. We also attended the first and second TeraGrid Conferences (June 2006 and June 2007), and weekly TeraGrid Architecture meetings. Finally, one of us (Zimmerman) served as co-chair of TeraGrid's Impact Requirements Analysis Team (Impact RAT). This team was charged to investigate, identify, and recommend ways to quantitatively and qualitatively measure TeraGrid's impact on scientific outcomes. Material from the group's report (TeraGrid Impact RAT, 2006) has been incorporated into this document where relevant.

3.1.5 Surveys

Three surveys were conducted over the course of the study. These consisted of a survey of a sample of current TeraGrid users and surveys to evaluate tutorials presented at the 2006 and 2007 TeraGrid conferences.

The purposes of the *TeraGrid User Survey* were to gain insight into the characteristics of those who use TeraGrid and to understand similarities and differences in the needs, motivations, and commitment of different types of TeraGrid users based on factors such as their experience with supercomputers, frequency of TeraGrid use, stage of career, field of research, gender, and age. Part 2 of this report contains detailed results from the user survey, which was administered using SurveyMonkey.com. The population from which the survey sample was drawn included all users who were active between 1 October 2005 and 30 September 2006 and all Principal Investigators associated with active projects. We stratified our population along two criteria: 1) the largest allocation associated with a user (e.g., DAC, MRAC, and LRAC), and 2) the field of science associated with projects. We selected a total of 595 individuals, representing a random, stratified sample proportional to the distribution of users by field and allocation category. The survey response rate was 52%.

We prepared and administered two 15-question surveys designed to evaluate TeraGrid education and outreach activities. These surveys were developed to measure attendees' satisfaction with pre-conference tutorials conducted as part of the TeraGrid '06 and TeraGrid '07 conferences. Findings from the survey of the 2006 TeraGrid conference tutorials are available in a separate report (Krause & Zimmerman, 2007). The results showed that the tutorial attendees who responded to the survey generally rated their experience positively and would attend another EOT event sponsored by TeraGrid. Responses to open-ended items suggested that some attention should be paid to the level of the tutorial and the ability of the attendees. Offering introductory tutorials in which attendees are given practical step-by-step training in tandem with more advanced tutorials directed at experienced users could be beneficial in meeting the needs of a wide variety of users. Preliminary analysis of results from the survey administered in conjunction with the TeraGrid '07 Conference show that participants rated the tutorials highly.

3.1.6 Internal Systems and Documents

TeraGrid senior management and resource provider personnel at all levels were supportive of our study. In many cases they agreed to be interviewed and in other instances they provided access to information and people which greatly facilitated our research. For example, internal sources such as the TeraGrid central database, which includes information on all projects with a TeraGrid allocation, were made available to us. We monitored several project email lists, and we were provided with copies of TeraGrid proposals and findings from NSF reviews of TeraGrid. There were sources of information, however, to which we were not privy, including the weekly meeting of the TeraGrid management team.

3.1.7 Supplementary Information: The TeraGrid Planning Process

In late spring 2007, the UM-SI received a separate grant from NSF to facilitate a planning process whose goal was to provide information to help guide the future evolution of TeraGrid. One objective of the first phase of the process was to gather information about the future requirements (i.e., over the next five years) of current and potential users of TeraGrid. During summer 2007, researchers and staff at the UM-SI organized and conducted a series of three workshops to collect information on user requirements that could be used, along with findings from the evaluation study, to inform subsequent phases of the planning process. The first workshop, which focused on the needs of those developing TeraGrid Science Gateways and on the intended users of those gateways was held in June 2007. Seventeen of the then 21 science gateways were represented at the workshop. The second and third workshops were held in late August 2007. The goal of these workshops was to gain an understanding of the requirements of current and potential future users of TeraGrid. The user workshops brought together 35 individuals from ten research domains to address the question: Where could HPC take you and others in your field over the next 5 to 7 years? The findings from the planning process workshops (Lawrence & Zimmerman, 2007a; 2007b) supplement data collected in the evaluation study.

3.2 Data Analysis

The qualitative data consisted of transcripts and notes from interviews, documents and web sites related to TeraGrid and the TeraGrid Science Gateways, and our notes and observations.

We developed a coding scheme to analyze qualitative data based on our research questions, interview topics, and prior literature.

As described in the detailed survey reports, quantitative survey data were analyzed using the statistical package SPSS®. We calculated descriptive statistics for all variables and conducted tests of association between variables based on the appropriate method in specific instances. The responses to the open-ended survey questions were coded according to major categories that emerged from an analysis of the data.

4. TeraGrid History, Organization, and Operations

NSF has been making substantial investments in HPC for more than 20 years. Although it was not an objective of this investigation to document the history of the TeraGrid project or to analyze the wider contexts under which it developed and evolved, in the course of the project we learned that aspects of this background were important to understand the data we were collecting. Therefore, 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, third-party reports, and internal TeraGrid documents available through TeraGrid's wiki.[2] Additional insights into the history that emerged from interviews are discussed in our findings and are not covered here. As we will show in the results section, the context in which TeraGrid was created and the nature of its growth affected the TeraGrid virtual organization and others' perceptions about the purpose of TeraGrid.

4.1 History

TeraGrid grew out of a late 1990's focus on terascale computing and grid computing.
The project, as it currently exists, is the result of awards made through a series of NSF solicitations and Dear Colleague letters.

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). Like the PITAC document, the workshop reports emphasized the importance of the most capable computers to U.S. competitiveness in science and engineering research.

What is now known as the TeraGrid began in 2001 as a collaboration among four sites: Argonne National Laboratory (ANL), California Institute of Technology (Caltech), National Center for Supercomputing Applications (NCSA), and San Diego Supercomputer Center

[2] See http://www.teragridforum.org/mediawiki/index.php?title=Main_Page. Although the wiki is not intended to serve as TeraGrid's public Internet site, much of the information on the site is unrestricted.

(SDSC). This partnership 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, n.p.).

The DTF solicitation identified a computational grid as a way 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." 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.

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 that year to expand the capabilities of the original DTF sites and to include PSC's 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). Although they were listed as separate resource providers, Purdue and IU partnered on a proposal and received a joint award.[3] In 2004, TeraGrid entered full production, and in 2005 NSF's newly created Office of Cyberinfrastructure (OCI) extended its support with a $150 million set of awards for operation, user support and enhancement of the TeraGrid facility over the next 5 years. As noted previously, one of us attended the first quarterly meeting organized by TeraGrid following the terascale extension award.

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

4.2 Organization and Operations

When the evaluation study began TeraGrid was comprised of eight resource providers: IU, ORNL, NCSA, PSC, Purdue, SDSC, TACC, and UC/ANL. In June 2006, the National Center for Atmospheric Research (NCAR) became the ninth TeraGrid RP site. Figure 1 shows the TeraGrid partnership during the period of our investigation. The tenth and eleventh sites—the Louisiana Optical Network and the National Institute for Computational Sciences established by the University of Tennessee and Oak Ridge National Laboratory—joined the TeraGrid in fall 2007 after the completion of our study.

Figure 1: TeraGrid Resource Providers - June 2006

As stated in the *TeraGrid Policy Management Framework*, the TeraGrid comprises "multiple cross-referenced but independent awards to autonomous institutions" (Catlett, Goasguen, & Cobb, 2006, p. 2). Each resource provider is funded by NSF to operate and support resources. The Grid Infrastructure Group (GIG), funded through a separate grant to UC/ANL, is a distributed body that provides integration and coordination across the project (Catlett et al., 2008). The GIG supports common services such as user support and authentication services and common processes such as accounting, authorizations, and allocations peer review. It also supports the TeraGrid operations and helpdesk, education, outreach and training activities, external affairs, software coordination and system architecture and planning, and a dedicated optical network backbone. Each RP selects a site lead to make and ensure commitments on behalf of the organization.

The TeraGrid Forum includes one representative from each institution (i.e., each RP site and the GIG); it is the main decision making body for the TeraGrid. Working Groups and Requirements Analysis Teams (RATs) help provide coordination across the project by addressing areas of common concern. Working Groups last for a year or more and generally include representation from all sites. Their goal is to coordinate services in major areas of the project such as accounting, allocations, data, user services, and security. Requirements Analysis Teams bring together a small number of experts to work together over a period of six to ten weeks to understand a specific topic or challenge, investigate solutions, and

recommend a course of action. The Cyberinfrastructure User Advisory Committee (CUAC), which along with the GIG and resource providers was mandated by NSF, was an advisory body comprised of users of cyberinfrastructure.[4] More detailed information about TeraGrid's organizational structure, governance, and operations is available in Catlett et al. (2008) and Catlett, Goasguen, and Cobb (2006).

Due to the distributed nature of TeraGrid, most communication occurs through electronic mail, over the Access Grid, and via teleconferences. The purpose of the TeraGrid wiki is to serve as a central point for sharing documents, information, and ideas. The RP site leads and the GIG meet face-to-face on a quarterly basis, and the annual TeraGrid Conference provides an opportunity for face-to-face meetings of working groups, RATs, and other teams.

Although this report focuses on an evaluation study of the TeraGrid, it is important to note that TeraGrid is only one activity and one source of funding for the resource providers. The percentage of each RP's budget and staff that is supported by the TeraGrid award varies across sites. In addition, TeraGrid is only one program that has, over time, supported HPC in the United States. For example, from 2001-2004 TeraGrid co-existed with the Partnership for Advanced Computational Infrastructure (PACI), which ran from 1997-2004. There were two PACIs—the Alliance and the National Partnership for Advanced Computational Infrastructure (NPACI). In summary, TeraGrid grew from and exists within a broader context. The ways in which past and present collaborations and competition among resource providers, changing user needs and requirements, and technical developments affect the TeraGrid project are analyzed in the results section of this report.

4.2.1 Access to TeraGrid

Similar to other HPC resources, an allocation is required in order to use TeraGrid. As described on the TeraGrid web site, the following steps are necessary to obtain an allocation.[5]

> To use TeraGrid resources, you must submit a request for an allocation of computing time or data storage space. To make such a request, you need to have an understanding of the type of codes you will be running or the amount and type of data storage you will need, the amount of time you'll need to complete the simulations you plan to conduct, and any special data needs that accompany a computing time request. Allocation requests are subject to a review process, which varies according to the size of your request.

As noted earlier, proposals are typically directed toward one of three award categories: DAC, MRAC, or LRAC. The awards differ based on the number of service units allotted. Service units are generally defined as "equivalent to either one CPU-hour, or one wall-clock-hour on

[4] In January 2008, the CUAC was reorganized into the TeraGrid Science Advisory Board (SAB). The charge to the SAB is to "provide advice to the TeraGrid Forum and the NSF TeraGrid Program Officer on a wide spectrum of scientific and technical activities within or involving the TeraGrid." See http://www.teragridforum.org/mediawiki/index.php?title=Science_Advisory_Board

[5] See http://www.teragrid.org/userinfo/access/allocations.php. TeraGrid began requiring allocations for data storage shortly after the completion of our study.

one CPU, of the system of interest," although exact definitions vary based on resource platform according to the NSF Cyberinfrastructure Resource Allocations Policy document.[6] All proposals have a PI. If an award is made based on the proposal, a project is established, and the PI may add users may to the project. Authentication means that a user has a TeraGrid account; authorization defines what that user is allowed to do at particular sites. For example, a user may be restricted to file transfers at SDSC, but authorized to access compute resources at PSC.

5. Introduction to the Findings

The findings of the evaluation research study are organized by the major categories of the evaluation as well as themes that emerged from the investigation. The end of each findings section (i.e., sections 6-9) contains a brief summary and some discussion of the implications of the results. As noted in the introduction, the latter should be viewed as preliminary and are provided for the purpose of giving some inital feedback on the results.

In addition to Part 2 of this report, which analyzes results from the *TeraGrid User Survey*, we have published other documents that present findings from specific data collection activities (Zimmerman & Finholt, 2006; Zimmerman, 2007; Krause & Zimmerman, 2007; Zimmerman & Finholt, 2007). Results and conclusions from these publications are incorporated into the relevant portion of the results sections that follow.

These findings reflect data collected at a particular point in time. For example, in summer 2006 TeraGrid personnel were debating how to define a TeraGrid user. This was also the time of the problematic implementation of version 3 of the Common TeraGrid Software Stack (CTSS). Although specific challenges will change over time, we use these events to illustrate the types of issues faced by TeraGrid participants.

6. The TeraGrid Collaboration

The TeraGrid collaboration can be characterized as a series of tensions that both create challenges and help the organization to grow, innovate, and adapt.[7] In this section we analyze the sources and affects of those tensions. One set of stresses arises from the diverse characteristics, histories, and cultures of the participating sites, their need to collaborate as well as compete, and their desire to retain individual autonomy and identity while cooperating under the TeraGrid banner. In addition, as new institutions join the TeraGrid scaling becomes a challenge. For instance, communication and coordination, management and governance, and technical integration become more complex. New participants must be "caught up" with the existing policies, procedures, and plans and the rationales behind them,

[6] See http://www.ci-partnership.org/Allocations/allocationspolicy.html

[7] Informally, we have heard some objections to the use of the word *collaboration* to describe the relationship among the resource providers. This opinion was expressed by individuals who view a collaboration as a partnership among individuals or organizations who willingly choose to work with each other. We use the broader *Oxford English Dictionary* definition, which states that to collaborate is "to work in conjunction with others."

and the sites must adapt to a reconfigured organizational arrangement. The concept of a *virtual organization* provides a framework to better understand these issues.

A second source of tensions stems from TeraGrid's three-pronged mission to support the most advanced computational science in various domains, to empower new communities of users, and to provide resources and services that can be extended to the broader cyberinfrastructure. Each of these goals can require different methods, resources, expertise, and strategies to achieve, making it difficult to establish and balance priorities and to define success. Further, some goals fit better with the existing strengths, customer bases, and priorities of individual resource providers than others.

Finally, there is a tension between the desire for infrastructural reliability and stability, particularly on the part of users, and the research and development that are often necessary to create a distributed cyberinfrastructure. In TeraGrid, achieving this balance has been particularly difficult because few users demanded the grid computing capability that TeraGrid was developing.

Below we analyze each source of tension, while also showing that they do not operate in isolation from each other. We find that much of what we learned in the study was knowledge that already exists in the project, but for various reasons it was not brought to the foreground, discussed, and acted upon. The results presented in this section are based primarily on interviews with TeraGrid personnel. Data collected during participant observations and interviews with others, especially cyberinfrastructure experts, also informed the findings discussed in this part of the report.

6.1 TeraGrid as a Virtual Organization

TeraGrid is a virtual organization (VO). As such it has the potential to capitalize on the advantages of organizing in this way, but it also confronts many of the challenges faced by other distributed organizations. DeSanctis and Monge (1999) defined a virtual organization as

> …a collection of geographically, functionally and/or culturally diverse entities that are linked by electronic forms of communication and rely on lateral, dynamic relationships for coordination (p. 693).

Virtual organizations vary across dimensions such as size, degree of formality, lifespan, and purpose. The definition of a VO is still evolving because it is a relatively new form and knowledge about it is incomplete. However, similar to DeSanctis and Monge, Cummings and his colleagues (2008) identified the following as being common to most VOs:

- *Distributed across space*, with participants spanning locales and institutions
- *Distributed across time*, with asynchronous as well as synchronous interactions
- *Dynamic structures and processes* at every stage of their lifecycle, from initiation to termination
- *Computationally enabled* via collaboration support systems

- *Computationally enhanced*, with simulations, databases, and analytic services which interact with human participants and are integral to the operation of the organization.

Cummings and his coauthors stated that technology use in a VO goes beyond communication. For example, a centralized project management plan that is accessible to everyone in the VO can help to coordinate work across sites by listing tasks, responsibilities, and milestones. The DeSanctis and Monge definition emphasized the lateral nature of relationships in a VO. Taken together, these two conceptual definitions provide a framework for understanding the TeraGrid partnership.

The existing literature has identified a number of reasons for the formation of virtual organizations. Among these are to facilitate access to resources and expertise, enhance ability to solve problems, improve economic and scientific competitiveness, help leverage limited funding and other resources, and enable discovery and innovation. But it seems that for every potential advantage there are risks.

> For example, greater geographical reach of the firm might be enabled via electronic communication, but the firm may also struggle with maintaining a coherent identity. Similarly, more participation by discussion in larger groups of people may be possible, but information overload may be a burden to participants; and more efficient communication might be possible but so might greater alienation (DeSanctis & Monge, 1999, p. 694).

Virtual organizations also face challenges related to building trust, motivating and rewarding participants, sharing knowledge, and scaling organizational processes and governance. In the section that follows, we analyze how these and other issues identified in the VO literature affect TeraGrid.

6.2 Factors Affecting the TeraGrid Virtual Organization

Cummings and his co-authors (2008) stated that one of the most important questions to ask about a VO is how it came to be created. In the case of TeraGrid what seems more relevant is the contrast between the way the project began and the manner in which it evolved into its current form. As noted in the history section, the goal of the original partnership between ANL, Caltech, NCSA, and SDSC was to create a distributed system based on homogeneous clusters.[8] During the time from 2001-2007, the TeraGrid grew from four resource providers to eleven. The expansion occurred through what we refer to as "organization by solicitation." Except for an initial cooperative proposal submitted by IU and Purdue, the current TeraGrid partners did not choose each other nor was growth planned strategically by TeraGrid. Resource providers were added primarily based on awards made to sites by NSF; the NSF also made the award to the UC/ANL to operate the GIG. The addition of sites increased the technical and organizational complexity of the project, and new initiatives such as the

[8] Two interviewees noted that the project actually began with two sites—NCSA and SDSC—and about three months later it was expanded to include Argonne and Caltech. The project's history is obviously complex. Our goal was to understand the affect of major events on the partnership, particularly as related by the participants themselves.

TeraGrid Science Gateways program introduced more complications and further clouded the vision. Thus, like many other VOs, TeraGrid has been and continues to be dynamic. While the situation has sometimes been stressful for TeraGrid personnel and often confused users, it has also forced the individual sites as well as the TeraGrid as a whole to adapt and innovate and helped to forge collaborations that otherwise might not have been formed. In the paragraphs that follow, we analyze the dynamism of TeraGrid and the affects of changing conditions.

6.2.1 Homogeneity to Heterogeneity

The first instantiation of the TeraGrid, in 2001, consisted primarily of Itanium-processor based machines distributed across the original four RP sites. The homogenous clusters were designed to operate as a single distributed facility and were linked via a dedicated optical network. Grid software was employed to integrate the resources to create the appearance of a "virtual system," which was in fact resources controlled independently by the individual sites. The recommended use guidelines emphasized grid applications, but according to one of the TeraGrid participants, users did not have grid applications, and so early on few requests were made for time on these machines. This is not surprising given the fact that "the user community that TeraGrid started with was not based on grid computing users," as one project member noted.[9] "The people who we catered to were not people running small jobs on small clusters. We need people who will spend millions of hours on full machine jobs running on big machines." Nonetheless, the lack of requests for allocations on high-end resources was troubling and problematic. As a result, the recommended use guidelines were altered to allow users to request time on a single computing system.[10] This change was one source of what became an ongoing lack of clarity within and outside the project about TeraGrid's vision and goals. However, there was little time to address this issue before further changes occurred.

The addition of new sites to the TeraGrid, beginning with PSC in 2002, created another challenge for the struggling grid computing vision. As a project member noted, "One of the hallmarks of the DTF was this complete homogeneity. They were all bought at the same time and constructed in the same way. PSC came in and had a completely different system." Specifically, PSC's Bigben supercomputer had a much different architecture and its networking design and equipment also varied from the other sites. With the addition of Bigben, homogeneity began to give way to heterogeneity. This posed new challenges for TeraGrid, which were summarized by one of the participants.

> So, it really made them have to rethink how they were going to add sites. It made them have to take into consideration ideas and concepts that hadn't been core or even critical to the original proposal that they submitted. I think that was difficult. I think it was really eye opening, and I think it helps especially if you think of the TeraGrid and the ETF as a prototype for the cyberinfrastructure. But I think it was also a challenge because here you've submitted a proposal that's gotten funded by the NSF and all of a sudden you have to change it. So, you still have all those goals and objectives that went along with the original proposal that you submitted. Then, all of

[9] This statement was supported by our interviews with users and with cyberinfrastructure experts.
[10] In section 7, we analyze reasons for the lack of grid applications both then and now.

> a sudden you have these new goals and objectives that are a little bit more difficult to meet, and you have to figure out how to do that and still make progress on both fronts.

TeraGrid personnel held different opinions regarding the merits of trying to achieve and maintain system homogeneity and enable grid computing. According to the participants, different views on technical matters are not uncommon among computer scientists. There are often multiple ways to accomplish the same task and different people and institutions have varied preferences. In this case, most interviewees were skeptical about the viability of homogeneity because they felt it was not scalable, desirable, or even possible. For example, one person who questioned the wisdom of this approach said:

> You can create a homogeneous environment that is sort of the least good of any one system. These systems have positive quirks, too, that are the very reason that people want to use them. So, you have to sort of manage those similarities and differences in a way that's usable.

Another project member, whose institution was not one of the original four sites, viewed it as an *improbable* feat even with identical clusters and an *impossible* task once new RPs and architectures were introduced.

> What was quickly demonstrated was unless you have everybody install the same version of everything at the same time, and ideally they should like all be clones—the system administrators should all be clones of each other—and somebody blows the whistle, and they push the button exactly at the same time, you will rapidly get differences, even if you have the same system from the vendor point-of-view. That's what I saw. Then, of course, we started bringing in new architectures and new sites.

Another person felt that grid computing was not given a chance to succeed, and the addition of new architectures increased the complexity and difficulty of achieving the original vision.

> I thought this was NSF's big chance to promote grid computing. Could those systems have been brought up and efficiently used as a true grid and true multiple resource by a multitude of people immediately? Absolutely not. Would there be wasted potential computing cycles? Absolutely. What better way could you support and increase the probability of evolving these capabilities and these technologies if not in an environment where this was top priority? The only people that should have gotten onto those systems were the people who had either proven, could argue, or were willing to do new things to try the new technologies from the very beginning.

In this, as in some other situations, it was not unusual for TeraGrid participants to have different interpretations of the outcome of the same situation. In this case, a couple of staff we spoke with felt the original goal had merit and that it had achieved a result that other experts thought was impossible. As one of them said:

> Back in the early days of the project...you could actually build and execute on one machine, so that you could do dynamic resolution of libraries rather than static resolution. And you could take it from San Diego to NCSA to Argonne and to Caltech, and the crazy thing would find all the libraries it needed at run time. So, that meant that the environment was really set up correctly, and you could roam between these resources. And I don't know if that still works today. I have my doubts. But that's kind of cool and kind of amazing.

The issue is not that people have varied opinions, but that the lack of clarity regarding a major aspect of the project leads to confusion and can result in personnel working at cross purposes. Several interviewees thought this was the case in regard to the issue of homogeneity.

> I see huge evidence that people are still working towards the homogeneity. I'm never sure if that's because there is somebody or some group of people who really believe this, or if there are just a set of people who have been given a charter several years ago and are still working toward that charter.

While PSC's hardware posed a challenge, it was, as one interviewee stated, "bringing them onto the playing field at the same level." The addition of the next set of resource providers significantly increased the diversity of the project membership and the heterogeneity of the technical environment. The affects of these changes are the subject of the next section.

6.2.2 Organizational Growth and Diversity

The nine sites that comprised the TeraGrid during the period of our study varied along several dimensions, including history, culture, vision and mission, organizational context, and the number, type, and size of the computational resources contributed to TeraGrid. The latter also corresponds to the number of personnel a site has available as well as the depth and/or breadth of expertise it brings to TeraGrid. According to the participants we interviewed, this diversity has been both challenging and beneficial. We found that diversity in terms of culture, mission, and human and technical capabilities, along with past experiences among some of the sites, were particularly important as both sources of tension and innovation.

NCSA and SDSC were two of the four supercomputing centers funded by NSF in 1985, and they have been the leading-edge sites for academic researchers in the United States. PSC was established by NSF in 1986, but it was not part of the PACI program. According to one of the interviewees, the terascale computing system award was PSC's "reentry into the NSF after having been out of it for awhile." Although NCSA, SDSC, and PSC provide the full complement of services typical of dedicated supercomputing centers (see Graham, Snir, & Patterson, 2005, p. 174), PSC is significantly smaller in terms of the number of employees. It was clear from our interviews with TeraGrid personnel, users, and CI experts that each center is perceived as having unique strengths. PSC is known for focusing on high-end users and dedicating people to work closely with them; NCSA is seen as serving a broad range of users, and SDSC is recognized for data.

The other six resource providers are situated in a multitude of contexts. The Computation Institute was established in 2000 as a partnership between Argonne National Laboratory and the University of Chicago. It is the home of the TeraGrid GIG. ANL, along with ORNL, is a Department of Energy (DOE) research center. Purdue's participation in TeraGrid is headquartered in the Rosen Center for Advanced Computing, which is located within the Office of the Vice President for Information Technology at Purdue (ITaP). Similarly, the Research Technologies Division, which is part of Information Technology Services, is IU's home for TeraGrid activities. Purdue and IU are smaller in terms of hardware capability and number of staff, and they have less experience providing resources, services, and support at the national scale. NCAR is unique from the other RPs in that it provides tools and technologies to a targeted area, namely the atmospheric and Earth system science community. TACC is a research center at the University of Texas at Austin. The following quote captures the impact of this organizational diversity as described by several interviewees.

> So there's the multi institutional challenge—institutions who have and are competing as well as collaborating. Each of them has a very different culture. Some of them have been in this field, obviously, for 20 something years. Others are new to this idea of sharing resources and providing shared resources. There are, of course, some sites that have, say, a DOE background, which is a totally different thing than the NSF culture.

Together, these factors can lead to what one person described as "serious philosophical differences among the sites."

In addition to philosophical differences, it was evident from our observations, and to a lesser degree from the interviews, that some sites have more resources, particularly in the form of computational capacity, human expertise, and total number of people to leverage and contribute to TeraGrid than other sites do, and this is a source of some tension.

> I know that we've put in twice as much effort as we're funded, and I suspect the same is true of [name of RP] from talking to the folks out there.[11] ...it sort of creates more tension in that there's an expectation for things to happen on money that doesn't even belong to them. You try to do what you can, but we've got other obligations on that money… As always, there's never enough money to do everything you want to do, and there's a lot of expectations from a lot of people.

Regardless of their size, every site has a stake in matters that affect them, and so each RP wants a say in how things will be done. Yet, sites vary in terms of their degree of experience and expertise in particular areas. For example, both TeraGrid personnel and users, particularly those who compute at multiple sites, remarked on the varied quality of support

[11] Sites are not named in order to protect confidentiality.

available at different sites.[12] Further, smaller RPs must consider whether they can implement or commit to particular tasks such as installing and supporting new software and making future upgrades. As one person stated, "Compared to the other places, we're pretty small. ... So, the number of people we have to throw at this effort is very small." This diversity raises questions about whose opinion should prevail when the capabilities of sites are not equal. Resource providers with more experience and knowledge in a particular area can feel that considering the opinions of sites with less capability slows down the process of getting decisions made or tasks accomplished. And for all sites, doing things in a cooperative way can be more time-consuming, at least in the short-term. CTSS, the means by which TeraGrid delivers a common set of software capabilities to users, is an example of this.

> With CTSS, for example, on the one hand, it's sort of a pain in the butt for everyone involved. On the other hand, once things get stable, it's one of those things that truly sort of help the people who manage the systems because of the packaging effort that goes on. For example, they shouldn't have to worry about endless dorking around trying to get software configured and running. They should just be able to run a command and it all works. We're not quite there, yet, so people have to put in the time both ways, which doesn't feel real comfortable. But I think those things in the long run will help.

Challenges regarding CTSS were exacerbated by a problematic upgrade in mid-2006 from version 2 to version 3, which one TeraGrid participant described "as the CTSS fiasco." The situation frustrated users, and in some cases, significantly hindered their work. However, the event spurred some personnel to reanalyze the CTSS components and the way in which they are delivered. As we ended our study, TeraGrid had decided to package version 4 of the CTSS software into task-oriented kits.[13] Each kit is designed to implement a set of related capabilities based on the needs of users, such as running jobs, moving data, or computing remotely. Resource providers are required to implement the core integration kit, but they can choose whether or not to make other kits available.

The level of resource capability in terms of compute cycles available at sites is another source of tension. Later, we will show how quickly the sites with the most capable computational resources can shift, but in late summer 2006, one of the interviewees stated that NCSA, PSC, and SDSC provided more than ninety percent of the resources that are allocated through TeraGrid. In addition, NCSA and SDSC have more than 20 years of experience providing computational resources and support at a national scale. Further, as we discuss in a later section, users want to compute on resources with which they are familiar and that meet their technical requirements. Therefore, they can be reluctant or unable to compute on the resource available at sites such as IU or Purdue.

---

[12] Both TeraGrid personnel and users were hesitant to name these sites, and we did not press the matter. The point is that personnel and users *perceived* a difference among site capabilities in terms of support. It is also logical to expect that sites would have different abilities in this area.

[13] Version 4 of CTSS has now been implemented according to the TeraGrid web site: http://www.teragrid.org/userinfo/software/ctss.php.

> I agree that it's helpful to have these sites to explore these technologies, but let's think about it that way. Let's not think about it as production resource providers. Let's think about them as development partners evolving these technologies. Of course, they are driven by the fact that they have an RP award, and they have to show usage of their resources. So, we're often doing artificial things to try to bring users to them.

In summary, participants do not share the same notions about what it means to be a member of TeraGrid. In addition, the roles and responsibilities of individual sites are sometimes obscure to other providers. This is probably an outcome of the way in which TeraGrid grew (i.e. sites did not choose each other) and is funded (i.e. each site has a contract with NSF).

On the other hand, according to many interviewees, the "forced" collaboration has also produced benefits. Personnel commented on the fact that TeraGrid created a "fabric of collaboration that didn't exist before" by bringing people into contact with others that they would not have met otherwise. This was mentioned by many of the TeraGrid personnel as one of the project's most important successes.[14] Even individuals who were critical of the project overall agreed with this sentiment.

> If there was anything good about the TeraGrid, I'd put that high on the list. It's enabled us−me in particular−to get to know a set of people that are good people to know and people I would not have got to meet otherwise.

Individuals attributed the opportunity to work with those at other sites as helping them to grow both personally (e.g., learning to negotiate with others and work across distance) and professionally (e.g., improve local processes or the way they manage and configure their systems). In addition, the tension among the sites encouraged what one person described as "a necessary push out of the old."

> If we're going to take a step from the old traditional paradigm of supercomputing, you need projects like TeraGrid and OSG and even to a smaller extent things like PlanetLab that come forward and say, "Look, we need to figure out how to coordinate all this stuff." They've had very different ways of going about it. TeraGrid has been very funded, very formal in this activity. OSG is more grass roots. The same with things like PlanetLab. The Europeans and other continents have certainly done things along these lines, too. I think having NSF really push it in terms of TeraGrid **will be** a good thing.

Some credited the leadership "at smaller sites" for their role in this and acknowledged that they "have been an asset" to the project. Referring to the national centers, an interviewee at one of these institutions noted:

[14] A comprehensive analysis of participants' view regarding project successes appears in section 6.3.3.

> There's a lot of momentum, a lot of—I don't know what the right word is—a lot of reluctance to move off of where they've been for a long time. Some of the smaller sites might do a better job of grasping and dealing with it.

The positive outcomes of bringing such diverse sites together in one project were not foreseen by most participants at the start of the project, although some expressed their initial excitement at the opportunity to be involved in TeraGrid even though, for most, this feeling diminished over time. The benefits have not come without costs, however, particularly in terms of stress on individuals and tensions between institutions. In the next section, we analyze the challenges of managing and governing diverse, distributed sites and scaling communication and coordination processes to deal with organizational growth.

6.2.3 Communication, Coordination, Management and Governance

Cummings and his colleagues (2008) observed that changes in the composition of a virtual organization may surface entirely new processes. The short period of our study limited the amount and type of data we were able to collect on mechanisms for communication and coordination and management and governance, but based on interviews and observations, we saw evidence that these systems were affected by TeraGrid's growth. In addition, interviewees expressed concern about how these processes would scale as more sites are added in the future. Below, we analyze each of these areas in turn.

Communication and Coordination

TeraGrid uses a number of mechanisms to communicate and coordinate across the distributed organization.[15] These approaches include Working Groups and Requirements Analysis Teams, regular meetings of TeraGrid management using telephone conferencing and the Access Grid, the quarterly face-to-face meeting, in-person gatherings at the yearly TeraGrid conference and other events such as the annual supercomputing conference, and numerous electronic mail lists. Centralized information systems such as the TeraGrid central database and the TeraGrid wiki also play a useful role in information sharing and appear to aid coordination.

As we and others have learned from prior studies of distributed projects, it is difficult to achieve an appropriate balance in terms of the amount and type of information shared. TeraGrid personnel who had participated in other collaborations noted this, too: "Anytime you have multiple organizations, you can never underestimate the overhead that you have to spend on communication, level setting, and getting the minds to meet." A problem we have witnessed in other investigations is what we refer to as "death by email attachment." We use this phrase to characterize the information overload that virtual organization participants often experience. This is a challenge for TeraGrid as well.

> There’s an enormous amount of time that can be spent sitting on these working group calls, following up with the email. I got very frustrated one day, and I started what I knew was an ill fated effort and so eventually I quit, but I was going to keep track of

[15] Another important topic for cyberinfrastructure projects is communication with user communities and with other external stakeholders (Spencer et al., 2007, p. 21).

> all the emails that come out that say, "Here's a document that you have to read and get back at the end of the week." What speed reading and comprehension rate did you have to have just to keep up—let alone if you could provide any input?

Those who needed responses to requests for information, concurrence on plans, and other feedback from individuals at remote sites expressed frustration at the lack of reaction to their queries. We witnessed a number of occasions where people continually requested replies to their calls for information, feedback on proposed actions, and input on decisions. Certainly, information overload contributed to low response rates, but we also observed that there were often unresolved issues that kept sites from responding. This usually became apparent in face-to-face and Access Grid meetings when individuals made comments or raised questions about the topic at hand, which indicated that the lack of response was due to confusion, disagreement, or reluctance to act upon something. Some people dealt with this problem by stating that they would move forward by a particular date if no input was received. This was not always a workable solution, however, particularly in cases where action was required by personnel at other RP sites in order for a task or activity to occur. Any action on site systems such as software installation, for instance, depends on direct participation from resource providers. As we shall see later, the inability to "force" sites to comply in this way is a source of ongoing tension.

The TeraGrid wiki was initiated in September 2006 and began somewhat as an experiment to test how it would work as a means for project personnel to share experiences, ideas, and information. TeraGrid had used technologies prior to the wiki such as a forum, so shared project spaces were not new. Although we did not collect quantitative data on its use by TeraGrid personnel or people external to the project, we observed that the wiki quickly became a rich resource for information about the TeraGrid. The wiki includes meeting agendas, notes and reports from working groups and RATs, presentation slides, and procedures and policies. It is now a routine matter for agendas, documents, slides, web links, etc. to be posted to the site in advance of or during meetings. The wiki also provides a significant view into project activities. Transparency is an important part of governance, and the wiki helps provide this for both internal and external parties.

While the wiki is a shared and open space for information, it only contains what is posted there, and it is not a total solution for communication and coordination in such a complex project. We observed and personnel commented on the fact that important meetings and decisions often go undocumented and that additional centralized information sources would be useful. For example, at the TeraGrid quarterly meetings we attended notes were rarely taken, action items recorded, or decisions documented,[16] and the weekly Architecture team meetings were documented irregularly. One of the project participants outlined the reasons why this type of documentation is important.

> There's a lot of policy missing that just doesn't exist—other than there were discussions that were had on these AG sessions we have on Thursdays—never

[16] The exception is votes on TeraGrid policy documents, which are recorded in the final version of the policy.

> documented. And so everybody walks away with their own understanding and goes and works based on that understanding. Well, it isn't a shared understanding, and so we run into various issues along the way where there is disagreement, and they all point to the same conversation saying, "This is what was said," and, of course, the interpretations are different. We lack a lot of structure in that sense. Now that is a fair amount of overhead that nobody ever wants to do. So, we suffer through anyway.

One result is that discussions are continually repeated if they are not documented. As one interviewee observed:

> It's a fine line to walk between too much process such that it becomes burdensome, but on the other hand you don't want to keep repeating the same debates over and over again because you haven't recorded the outcomes of the decision from these things…so that when something gets decided, it can be put behind and you can move on and not have the debate again in 6 months.

Similarly, another person noted that when decisions are not recorded it is difficult to discern if an issue was dropped because people got too busy to work on it, or if there was a strategic decision made to not do something.

Besides the recording of decisions and meeting notes, some TeraGrid personnel expressed a desire for additional centralized sources of information. The TeraGrid central database was held up as an example of a useful repository for information on users and allocations. For example, helpdesk personnel query it to learn more about individuals who contact them with problems (e.g., where users have accounts and projects they are associated with). But by itself, the central database is inadequate. For example, someone noted that there are "pieces that are commonly handled at the centers that cannot be handled by the central database" such as refunds. Another interviewee spoke to a broader need for shared sources of information.

> There are things like a common good that you would think the government would provide, for example, things like databases of central information. … We've needed a central database for machine downtime information for awhile, and there's not really a resource for that kind of thing.

This person explained that a real-time database with information on system downtime would be useful to several internal groups as well as to users, who could use it to compare resources and decide where to submit their jobs. The interviewee noted, however, that one hindrance to this goal is the difficulty of getting a person at each RP to commit to update the information for their site; this is a challenge that grows as the number of sites increases. Another potential impediment to central data sources is that they reveal information that sites may prefer not to share or make known. There are data that indicate this may be a consideration for some TeraGrid personnel. For example, the Inca test harness and monitoring project allows TeraGrid to run tests, benchmarks, or any script or executable on all machines, collect the information centrally, and display it from one location. It is used to monitor CTSS and, at the time of the evaluation study, it was being explored for measuring the performance of

GridFTP and client software. Two interviewees noted that members of their staff perceived Inca as an imposition rather than a tool to help them get their work done or a means to meet the goals of their site as well as those of TeraGrid.

> That has created some tension because I think for some of the sites—maybe all of the sites—perceive that as sort of a compliance thing. And I viewed it that way for a long time, too. Where it was the GIG trying to say to a site, "You're doing well. You're doing poorly." … We want to get to a point where we're monitoring the critical things that we need to monitor, and we're able to get the sites to respond to it without feeling like they're being undercut.

Many of the issues we discuss in our analysis of TeraGrid as a virtual organization are intertwined with each other. For example, if trust is lacking, people are apt to be suspicious of systems that monitor performance rather than to view them as aids to help them achieve their goals. This is even more likely to be the case when the systems require extra effort to implement or maintain or when procedures differ from local practices. Further, circumstances do not remain static. For example, coordination challenges have varied at different points in the project. Contrasting initial versus later stages of the project, one person noted:

> When you're trying to build something it's pretty clear what you're trying to accomplish, especially when it's hardware and software. You can say, "We're going to specify a system, procure a system, install it, and get it operating." There are some clear milestones that can be achieved. You can clearly see the results of those efforts. You can do it across organizations. "[Name of RP] got their hardware installed. Check, check." During an operational phase there's a lot of work where pretty much no milestone is achieved. ... Sometimes there's no clear goals because what you're doing is cutting edge and research-oriented.

Finally, most interviewees, even those who were positive about the current state of affairs in TeraGrid, were concerned about the ability to scale processes as the number of sites increase.

> There are a lot of good people out there that are capable and willing to contribute and interested in contributing. It’s efficiently managing and coordinating the work of all of these people that is extremely difficult.

Recording decisions, having clear processes in place, and holding people accountable are challenges that many TeraGrid personnel feel are currently not dealt with as well as they could and need to be handled. Capturing and sharing the "right" amount and type of information is challenging as is scaling mechanisms as the organization grows and changes. This mirrors findings by Cummings and Kiesler (2005) who observed that the number of institutions had more of a negative impact on a distributed project's ability to meet its desired outcomes than the number of disciplines involved. Further, they found that as more institutions became involved collaborations were less likely to employ coordination activities that might help them mediate such challenges (Cummings & Kiesler, 2007). Organizational growth poses problems for management and governance structures, too.

Management and Governance

Communication and coordination mechanisms play an important role in management and governance. For example, as stated above, transparency and the involvement of key stakeholders are two necessary conditions for effective governance. Although it is not a complete solution, the TeraGrid wiki appears to support these goals in effective ways. In this section, though, we turn attention more directly to management and governance.[17] Specifically, we analyze the views of TeraGrid personnel regarding project management and governance. Again, we see a set of tensions at work, primarily between attempts to balance autonomy and interdependence, local and project-wide demands, and the present and future needs of users.

The Institute on Governance defined governance as "a set of ideas about how direction is provided to human activity."[18] 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. Management in TeraGrid relies heavily on a matrix approach. In many cases, individuals who are responsible for particular areas or tasks are not the direct supervisor of those who are working on those areas. Further, their "employees" may not be located at their institution and may, in fact, be spread across several sites.

> In the GIG, we fund—partially—an awful lot of people, but it's sort of like the worst form of matrix management. We don't necessarily have contact with their managers. We don't even necessarily have that much day-to-day contact with the individuals. We do some tasking, and we do follow-up, and we try to keep people engaged. For people who are funded off the GIG and who are leading some of the working groups and are driving some of the major efforts forward, that's a little bit easier. But it isn't always clear that the people who are GIG-funded know they are GIG funded; know what they are trying to accomplish with the GIG funding, even if they've been told, and in the vast majority of cases, there's no contact between the GIG and the managers of the people, who are funded. So, those people live in the space of conflicting priorities. And how that plays out depends on the site cultures and everything else.

This quote emphasizes points made by other interviewees. First, most people have a fraction of their time funded by TeraGrid. Studies of other VOs have shown that level of effort on the collaborative project plays a role in determining how engaged people are or if they even are aware that they are part of the project (Lee, Dourish, & Mark, 2006; Olson et al., 2008). TeraGrid recognized and made an effort to avoid this problem. For example, most of the Area Directors and GIG personnel are full-time on the project. Second, even if individuals are aware that they receive a portion of their salary support from TeraGrid, they can be torn between local and project-wide demands. As one person said, "People have different

[17] Earlier in this report (see section 4.2), we described TeraGrid's governance structure as it existed at the end of our investigation. We do not repeat this information here.
[18] For further information, see the web site of the Institute on Governance: http://www.iog.ca/

priorities. They have different tasks. They have different opinions. And that's part of what makes any distributed project as a whole difficult." A third point that the quote above illustrates is that when an individual is responsible for coordinating the work of people at multiple RPs, he or she must figure out how to best engage each site. One person described this task as "... finding the path of least resistance. Each institution has a sub-award with a PI. Some PIs are more engaged than others. It's just understanding how different institutions work." Different people used different tactics. Among the approaches employed were visits to the other sites, formal and informal face-to-face meetings at conferences, and the use of project management tools.

There was widespread agreement among interviewees that an authoritative, hierarchical approach to governance and management would be ineffective for several reasons. Most of these have already been discussed in other parts of this report. These reasons include:

- funding arrangements, which make it logically impossible for the GIG to mandate or enforce policies;
- differences in technical environments across the sites, which can render it hard to implement standards and configure systems identically;
- the needs and demands of an RP's user communities, which can conflict with TeraGrid activities, particularly if those activities impact system stability; and
- diversity in institutional culture, mission, and human and technical resources.

Interviewees emphasized that maintaining a degree of autonomy increased in importance as the organization grew.

> I'm sure you're familiar with how this whole thing started with DTF and three identical machines and all the concepts were built from there. And when you start expanding to different architectures, different people, wider ranges of hardware and software things have to change from a very rigid structure to finding somewhere in between the very rigid and the completely autonomous structure that works; this is difficult.

The above quote also shows that in spite of consensus on the need for each site to maintain some autonomy, it is a constant struggle to find the appropriate balance between that and interdependence. Further, most personnel agree that some areas demand decisions that have "more teeth," as one person described it at a meeting, but it is difficult for the group to reach consensus on the matters that require standardization.

> You want to proactively catch problems before the users notice them. I really think that's where we need to get to. The trick is to figure out how to do that, but at the same time, respect the autonomy of the RPs. There are things we're trying to do together, but they are independent sites with independent goals and their own user communities and their own things that they are trying to protect and foster and nurture and everything else.

The processes by which decisions get made, direction is set, and the project is managed are important in a distributed collaboration, but leadership was mentioned as being critical, too. This was a common sentiment, but beyond that, interviewees spoke to this topic in various ways. In general, they were more reticent on this than on other subjects.[19] Some TeraGrid personnel were bothered by the fact that the site of the GIG was selected by NSF, even though most of the people who mentioned this were not openly critical of the individuals who comprised this body. In fact, some members of the GIG were highly regarded by others in the project. In other cases, interviewees acknowledged that the GIG, particularly Charlie Catlett, who was GIG director during the course of our study, faced many obstacles. These arose from the management and governance challenges we discussed above as well as from demands and restrictions that were perceived as coming from NSF and from a lack of engagement and support for the project from leadership at some of the RP sites. A few interviewees, however, spoke more directly to the issue of leadership.

> Although I talked about the challenges are really people, what I see happening with this project is that it tends to want to move and become more about solving technical issues because you're dealing with technical people. When you have an environment like that, and you don't have the right kind of leadership, technical people are going to do what they're very good at; they're going to close the door, and they're going to pound away on code. It's a natural thing to happen, but you have to counter that with very strong leadership and vision that keeps people pulling back to the overall mission. You continue to work on technical challenges, but that is not the reason TeraGrid is being built. There's a bigger reason for this. And that's what I'd like to see. I'd like to see a little more visionary leadership.

It is important to note that interviewees talked about the importance of leadership beyond the GIG director. A project as large and complex project as TeraGrid requires leadership on multiple fronts and from more than one person. Speaking of TeraGrid working groups, one person said, "Give the group some direction. Otherwise, I think you tend to find that you've got energy and you've got expertise, and they want to do something." Leaders emerged in the course of the project and achieved the respect of personnel at remote sites by stepping forward and accomplishing tasks that needed to be done and carrying out the work in a way that was respectful of the needs and situations of others. Some people in positions of leaderships were identified as being effective in their role while others were not.

It was clear from both our interviews and observations that many people involved in TeraGrid want it to succeed and believe the project could accomplish important work that would benefit users. There was also frustration around the issues that made success slower, more difficult, or stressful to achieve—leading, governing, and managing multiple and diverse organizations were part of this challenge.

[19] We did not ask TeraGrid personnel direct questions about management, governance, or leadership. These were areas that interviewees identified, particularly in response to our questions about project challenges and successes.

6.2.4 Collaboration and Competition

The last source of tension we analyze in the TeraGrid virtual organization is the one between collaboration and competition. There was widespread, although not total, agreement among interviewees in all categories that the continual demand for RPs, particularly NCSA, SDSC, and PSC to collaborate as well as compete was detrimental to TeraGrid and to its users. While some individuals acknowledged that competition could spur innovation, most felt that the current situation was out of balance.

> I think the other big challenge is… What's the term for it? Coopertition. NSF exasperates this problem because they're making the competition cycles shorter, but then they're expecting everyone to work together to make this a success, but "By the way you have to work together and show this to be a success, but then you are going to be competing." ... So, from my perspective there's a constant tension between: Well, do you do this and make TeraGrid succeed knowing, for example, at the end this is going to make University of Chicago/Argonne look better because they'll take the glory as the GIG leaders, or do you do something that maybe isn't as good for TeraGrid that makes better sense for [name of RP], or do you try to do the logical thing which may be half way in between these two things? And there are constant tensions on that front. It's not that people don't try to do the right thing, but the question is, "Whose right thing are you going to try to do?"

Short funding cycles were seen as exacerbating "coopertition" as well as being an ineffective means to support persistent, stable, and reliable infrastructure. In regard to this, one of the CI experts stated, "You don't build infrastructure in three year chunks. They have to be ten years." He added, "NSF needs to get out of the mind of recompeting these things. These are national facilities. They need to be managed as a facility and not a single investigator proposal."

As noted earlier, interviewees emphasized the competition/collaboration tension as a particular problem for NCSA, PSC, and SDSC, but this could be an artifact based on the timing of our study. The circle of competitors for NSF HPC solicitations is growing. A prime example of this is NSF's Track 2 initiative, a four-year program to fund up to four leading-edge computing systems that will be integrated into TeraGrid.[20] The first of the Track 2 awards was made in September 2006 when TACC received a $59 million dollar, 5-year award from NSF for what is currently the most powerful computational resource on the TeraGrid.[21] A second Track 2 award has since been made for a computer that will be housed at the National Institute for Computational Science. As a result of this award, NICS became part of TeraGrid. These grants include support for personnel and operating costs. One interviewee predicted that they would change some of the dynamics in the partnership.

> ...once there is a Track 2 award made I suspect that whoever that is, is going to play the 800-pound gorilla because they will, in fact, be 60% or more of the total set of

[20] "National Science Board Approves Funds for Petascale Computing Systems," NSF Press Release 07-95. See http://nsf.gov/news/news_summ.jsp?cntn_id=109850&org=NSF&from=news

[21] The system, named Ranger, was dedicated on February 22, 2008.

> resources. Look, if it was me, I'd say, "This is the way we're going to do it at [name of RP] because this is what we've got and the support to do with what we can do, and I suggest that others follow suit.

Similarly, another person noted that "the institution that wins the first Track 2 award is going to have a big bat to swing." Behind these statements are tones of distrust and competitiveness, but there is also recognition that these systems are specialized and unique and must be managed carefully. Although sites are frustrated when other RPs do not install specific software, for example, most interviewees recognized that these decisions can be complicated. In particular, individuals recognize that system reliability and stability are major concerns for the sites and their users. As one person said, "It's important for RPs to have some autonomy and be able to say, 'This is what works for my users.' Each RP knows their users." Even so, it is difficult to discern a shared notion of what is acceptable in this regard particularly because, as we noted earlier, there is often more than one way to achieve a technical objective, and people have different opinions regarding the best approach.

6.3 TeraGrid's Vision and Mission

What is TeraGrid? What are its vision and goals? What are the priorities among the three elements of its mission? To the majority of TeraGrid personnel we interviewed—as well as to many users—the answers to these questions were unclear. In the absence of clarity, project participants did their best to answer these questions for themselves. In this section, we analyze the factors that contributed to this confusion and the circumstances that have made it difficult to resolve. We have already discussed many of them, which we reiterate briefly here. However, we focus our attention on the opening questions for two main reasons. First, in the absence of shared vision, it was difficult for those both in and outside the project to evaluate successes TeraGrid had already achieved as well as to define what it would mean for the project to succeed ultimately. Second, we examine the conflicting and sometimes new demands that TeraGrid's multi-faceted mission places on the technology and on the resource providers. We find that TeraGrid personnel held similar views about the problems caused by the lack of a shared concept of success and the difficulty of balancing the three-part mission. Further, their visions for the project, as gleaned from their conceptions of success, often overlapped. Unfortunately, these commonalities were seldom brought to the fore and openly discussed or debated.

6.3.1 What *is* TeraGrid?

As early as the first and second quarterly meetings in 2005, TeraGrid personnel who attended the meetings recognized that most users had difficulty understanding what TeraGrid was and how they might use it. TeraGrid was very new at this point, and it seems natural that it would take time for the project to form an identity, particularly in light of the fact that few users had grid applications. However the question of what TeraGrid *is* has persisted.

When we asked TeraGrid personnel about the project's challenges and successes, the answers we received were often preceded by comments such as,"I guess it depends on what the goals of the project are," or, It rests on "this idea of vision and what it is we're actually trying to accomplish." In the absence of a shared understanding of these topics, individuals did their

best to form their own answers and to direct their energies in those directions. As a consequence, we found the situation to be largely as one person predicted.

> If you ask people, "What is TeraGrid? Or what are the goals of TeraGrid?" you're going to get a lot of different answers. I'm not sure that I could even answer what the goals of the TeraGrid are. There's sort of the very lofty goal to make more computing available for science or to make more science possible through computing, but that's right next to world hunger. That doesn't really help us get anywhere. I think in some sense we're still sort of struggling: What is the TeraGrid?

Similarly, another person said, "If you asked five people at different institutions: What does TeraGrid mean to you? You'd get a lot of different answers because we haven't formed it up yet." People tried to define TeraGrid both by what they perceived it was and what it was not. For example, "TeraGrid is a unifying software environment and a dedicated network that enables certain things for users," but "traditional single resource supercomputing is not something TeraGrid enables."

Another problematic question that interviewees identified is: What is a TeraGrid user? This was a matter of some controversy during summer 2006. At that time, if a TeraGrid user was defined as anyone using a resource available on the TeraGrid then the number of TeraGrid users was several thousand.[22] If the definition was limited to users employing grid and middleware software, the number was very small. Most interviewees who addressed this topic did not believe that those who used the resources in "traditional" ways should be counted as TeraGrid users.[23] For one, it made it difficult to distinguish the contributions of individual sites from those of the TeraGrid. Further, interviewees perceived that for most users TeraGrid is "just a new name for the resources they use to get their science done."

> It sort of harkens to a discussion that happened last week…trying to provide metrics for TeraGrid. They were showing all TeraGrid users. As of March, April of this year, the number doubled. Why did the number double? Because we decided to call all the resources at NCSA and San Diego that weren't already TeraGrid resources, to call them TeraGrid resources. So, their users were called TeraGrid users. The users don't care; they didn't even know for the most part, except it was an inconvenience for them to have to change all of their projects and allocations, but other than that, it's no different for them. But all of a sudden, we have twice as many TeraGrid users. So, I was asking this question, "What do you mean by a TeraGrid user?" And there's no good definition of what that means. I say it's actually a meaningless term because there are very few users who care about TeraGrid in the context of what TeraGrid wants to be.

[22] This appears to be the definition that TeraGrid ultimately settled upon.
[23] There is not a strict definition of this type of use, but generally it was described as batch use of a single resource at one or more sites.

Interviewees noted that it has been difficult, almost from the outset, to define the goals of the project and that this continues to be something they wrestle with. The quote below is typical of what many said.

> What did the users need, what community were we trying to serve, and did the traditional supercomputing users really want to do grid computing? How could we make it attractive to them? How can we provide facilities that they would find useful? I think we've continued to struggle with those challenges.

The same questions held true for TeraGrid Science Gateways, which we discuss later in this report (see Section 9). Science gateways serve as front-end interfaces to TeraGrid resources. They attempt to hide the complexity of using TeraGrid and are a significant part of TeraGrid's strategy to reach new communities of users. Interviewees noted that the number of users was one of the few quantifiable measures that could serve as an indicator of success. Thus, there was competition for users between TeraGrid and the RP sites and, to some extent, between TeraGrid and the science gateways.[24] Like other large-scale, multi-million dollar projects, TeraGrid was continually under pressure to show results, but most participants believed that the "user count issue is a red herring."

> It's not so much that we've got 4,000 users or 4 million users, or we've got supercomputers with so many petaflops. Those are all things you need, but it really comes down to the researcher and those that use the computers and working with them to make sure that they're really getting the potential out of the resources that we provide.

Not all personnel were confused about the meaning of TeraGrid. A minority of those we interviewed seemed clear about the project's goals and for the most part, did not question them—at least in their conversations with us. Another few felt they understood what TeraGrid *is*, but they disagreed with what they perceived as its primary aims. For example, it was common for TeraGrid personnel to talk about the same activity or goal, but to have different views regarding whether it was a worthy one to aim for and whether it had or was likely to succeed. Examples of this include CTSS, roaming accounts, and the global parallel file system.

We found that the notion of what TeraGrid *is* represented uncertainty around three specific areas: TeraGrid's *vision*, which is related to what it means for it the project to be *successful*, and priorities among the three facets of its *mission*. We identified multiple reasons for ambiguity about these topics based on the data we collected.[25] Early factors, which we discussed previously, included the move from homogeneous to heterogeneous architecture,

[24] At the time, most science gateways had few users. Further, the users that did exist were not heavy users of TeraGrid resources available through the gateway.

[25] We cannot assess if the vision, mission, and goals of the project that included ANL, Caltech, NCSA, and SDSC were clear to those participants, at that time. We can only say that interviewees felt that in contrast to TeraGrid, the purpose of the DTF was more well-defined. This is not to say that everyone agreed with the goals of the DTF, as we discussed earlier in this report. In addition, the time-frame of the DTF may have been too short for ambiguity to emerge.

the relative absence of users for grid computing, a lack of consensus on the definition of grid computing, and the solicitation-driven addition of multiple RP sites over a short period of time. All of these factors increased the social and technical challenges of integrating distributed resources using a suite of grid and middleware software. In addition, we also analyzed challenges that stem from TeraGrid as a virtual organization such as "coopertition," lack of trust, and pressure to produce results quickly and continuously. All these aspects made it difficult to have the discussions that might have helped participants achieve clarity. Other reasons mentioned by interviewees included a lack of strong leadership, mixed messages from NSF, high expectations to deliver on many different tasks without a sense of present and future priorities, and the difficulty of architecting competing requirements on top of existing infrastructure.

6.3.2 Vision

There are numerous strategic planning sources that are intended to help an organization identify what *it* is, what it *does*, and *how* it does it. Vision and mission statements are important aspects of most strategic plans as are general statements—often called goals—outlining the ways in which the mission will be accomplished. The mission describes the overall purpose of an organization, including why it exists, its business, and its values. The vision is broader than the mission; it tries to answer the question of what success will look like when an organization is working optimally in relation to its environment and its key stakeholders (Bryson, 1995). Thus, vision and success are intertwined. It was difficult for most TeraGrid personnel to articulate the project's vision; it was easier for them to talk about success, particularly their views about what it would mean for the project to succeed in the long run. Although we will show later that there are several types of users, which created uncertainty about whose needs take precedence, TeraGrid personnel were unequivocal that the project's focus should be on enabling researchers to do their work.

TeraGrid personnel used the words mission, vision, and goals somewhat interchangeably. We are not concerned here with textbook definitions, but with the way in which TeraGrid participants and those external to the project viewed issues related to the project's vision, definition of success, and mission. We analyze each of these in turn, beginning with vision.

Until recently, it was difficult to find a vision statement for TeraGrid.[26] John Bryson, an expert on strategic planning stated that the absence of a written vision statement does not lead to failure nor does the presence of one guarantee success.

[26] We searched the TeraGrid wiki, TeraGrid public web site, the main text of annual reports covering October 2004 through calendar year 2006, and Catlett et al. (2008) for a vision statement. In August 2007, Dane Skow, who was then the GIG Director, presented the following vision during an overview of TeraGrid that he gave at a user workshop conducted as part of the TeraGrid planning process: "TeraGrid will create integrated, persistent, and pioneering computational resources that will significantly improve our nation’s ability and capacity to gain new insights into our most challenging research questions and societal problems. This vision requires an integrated approach to the scientific workflow including obtaining access, application development and execution, data analysis, collaboration and data management."

> While it may not be necessary to have a vision of success in order to improve organizational effectiveness, it is hard to imagine a truly high-performing organization that does not have at least an implicit and widely shared conception of what success looks like and how it might be achieved (1995, p. 156)

Bryson also stated that a vision statement is more difficult to develop than a mission because it "must usually be a treaty negotiated among rival coalitions" (p. 154). The majority of TeraGrid personnel we interviewed felt that the project lacked a shared vision, and they perceived this as a significant hindrance to planning, decision-making, and the ability to meet user needs. It was clear from our interviews and observations that TeraGrid personnel are dedicated to enabling science and to supporting users in that quest. Thus, low interest from users in the capabilities and functions that TeraGrid was developing caused many personnel to question the relevance of the project's initial aims or what they perceived as its present goals. This was compounded by the fact that many TeraGrid participants felt that users had other, and more important, needs that were not being met.

> At some level, you start with an elegant vision from the computer science point of view. You ultimately fall into the, "If you build it, they will come," but they don't. Because to you as the person who builds this, this is your project, and it's cool and you like it, and CTSS, Globus, whatever; it's cool, it's elegant, it sounds great and you map it out. That's your job. But to the physicists, they're thinking, "What am I going to publish? How am I going to keep my PhDs going and my post docs fed?" The preoccupations are just so different, and they're not with, "How can I use this cool tool that the TeraGrid is providing me?"

Since a vision is future-oriented, there is a natural stress between the current state and the desired future. Bryson described this as a useful tension, but he noted that goals "must be set high enough to provide a challenge but not so high as to induce paralysis, hopelessness, or too much stress" (p. 158). As the quote below illustrates, balancing the tension between giving users what they want today and readying them for the future is difficult.

> We're focusing on trying to put in all these grid capabilities that nobody cares about instead of trying to make it useful for the way people want to use things. Eventually, it'll evolve to that, but people aren't ready for that. And you need to push them, and it's important to do that. But I think we've gone too far in that end of it and not provided the functionality that they want today versus trying to push them toward tomorrow's functionality that they're not ready for yet anyway.

Overwhelmingly, TeraGrid personnel want to see the vision driven by user needs. As one interviewee put it, "I don't want some crazy Utopia interfering with the ability of the user to continue doing their science now as they see fit." At the same time, he and others believed that users need encouragement to try new things and that part of their job was to help users see the possible advantages of new capabilities.

> If you let users set the agenda, you're going to end up doing business as usual and maybe not making progress in terms of advancing the technology. However, if you

> are just a slave in advancing the technology, you ignore the users. This is where I think the leadership of the project has to say what the balance is between those things.

It was difficult for TeraGrid personnel to reconcile their desire to serve users when so few users were ready or able to use what TeraGrid was constructed to provide. The disparity between what TeraGrid personnel perceived users as needing and where TeraGrid efforts were directed left many project participants frustrated and in search of a reconcilable vision. Some individuals thought that TeraGrid should be a prototype for cyberinfrastructure both technically and organizationally.

> What I would like to see happen is TeraGrid may not be so much a technical challenge as it is a way of moving the whole nation forward in research by combining the resources.

Others agreed, but they expressed uncertainty about whether that was the intended vision for the project.

> I don't know if TeraGrid is really a technical challenge or it's more a challenge as to how faculty and universities work together to pool their resources for the better.

Individuals outside TeraGrid expressed similar ideas. For example, one of the CI experts we interviewed described the major challenge of cyberinfrastructure as getting distributed elements—people, resources, and data—to work together as a whole. He noted that socioeconomic, behavioral, and technological challenges would have to be overcome for this to happen, and he viewed this as the most important aspect of TeraGrid. He also said:

> I'm not too concerned about whether broad elements of the scientific community are using the TeraGrid directly because I personally believe the way the TeraGrid will impact the life of the average scientist 10 years from now is through the technologies that were developed and disseminated and distributed widely. … TeraGrid is doing many pioneering things that will find applications beyond the TeraGrid and beyond science.

Interestingly, the CI expert's prediction has been one aspect of TeraGrid's success to date according to TeraGrid personnel and to some users. In one sense, this should not be surprising. In spite of a lack of consensus regarding the definition of grid computing and the technical and organizational challenges resulting from the heterogeneity of RP resources, policies, and culture it is clear that tying together distributed resources and making it possible for users to do things across sites have been important goals since the start of the project. And it has remained an aim in the midst of unplanned growth in the number of RP sites. To accomplish this at the scale and complexity of TeraGrid required the development of new technologies. Sites also had to be willing to adapt local policies to meet project-wide needs. Unfortunately, without a shared notion of this as an important outcome, TeraGrid personnel and users do not know how to assess the value of what has been achieved in this regard.

6.3.3 Success

What success has TeraGrid achieved to date and what will define its success in the future? The TeraGrid personnel we interviewed identified three categories of success from the early stages of the project through the time of our study. These areas included collaboration across sites, development of new technologies and mechanisms to enable distributed computing, and broadening the base of TeraGrid users. Ultimate success for TeraGrid was defined in terms of the science that TeraGrid made possible. Some of the disappointment and frustration that TeraGrid personnel expressed about the project appears to be due to disparity between what individuals perceived as success so far versus ultimate success tied more directly to user needs and research advancement. Below, we analyze each area of success as described by TeraGrid personnel. In addition, we contrast the views of project participants with those of others', particularly users, as this comparison provides additional insight into the tension between TeraGrid efforts and activities at the time of our study and the needs of users as perceived by both TeraGrid personnel and users themselves.

Collaboration across Sites

Earlier, we reported that many of the TeraGrid personnel we interviewed stated that cross-site collaboration has been an unexpected and valuable outcome of the project. At the organizational level, individuals were surprised that resource providers were able to cooperate in such a competitive environment. They noted that the situation was far from perfect, but the fact that collaboration was possible at all astonished many. At the individual level, people enjoyed the opportunity to work with and to learn from personnel at other sites.

> I think from a people and staff wise perspective—compared to the independent supercomputing centers—we are closer tied. I know what the other centers have now. I know what their capabilities are. I know the people there. So, if I have a scientist with a need, I have a bigger picture of what can happen. So, I think one of the successes is the fact that all of the centers now cooperate a lot more than they did before. In spite of our growing pains, we are still getting a lot of science done.

Others' views about the collaboration between TeraGrid RPs were mixed. For example, the CI experts we interviewed expressed varied opinions. One person felt that collaboration was a major goal of the project and that it would occur with time and another believed that signs of increased cooperation among sites were already evident. On the other hand, two of the CI experts were critical of the ability of sites to cooperate; one person referenced cultural reasons and another believed that NSF's short funding cycles contributed to the tension between collaboration and competition among RPs. Some users, including science gateway developers were indirectly critical. In general, these individuals had less insight into or interest in the organizational dynamics between the sites. Their concerns were related to the way the various kinds of heterogeneity impacted their work. For example, science gateway developers mentioned the time required to work through similar issues with each site when a central point of contact would be more efficient for them (see also Lawrence & Zimmerman, 2007b). As one person remarked, "We only deal with four sites, and that's still a fair amount of overhead."

Technological Developments and Common Processes

The ability of sites to collaborate contributed to the development of technologies, processes, and procedures intended to make it easier for users to work at different sites. TeraGrid personnel noted several types of accomplishments in this category. For one, individuals expressed surprise that the system worked at all given the heterogeneity and complexity of the technical environment. After describing some of these challenges to us, one person said, "The fact that we're now dealing with the complexities that we discussed is a success."

Another and more frequently mentioned area of success was centralized processes such as those for accounting and allocations along with the centralized help desk for user support. Single sign-on login was also noted in this category.

Data transfer capabilities, particularly GridFTP and the high-speed network that facilitates data movement were also named by many TeraGrid personnel as technical achievements. Another data-related advancement that several TeraGrid interviewees identified is TeraGrid GPFS-WAN (Global Parallel File System-Wide Area Network). This is a 700 terabyte storage system mounted on several TeraGrid platforms. The system is physically located at SDSC, but it is accessible from all platforms on which it is mounted and appears to the user as a local directory. When asked about the ways that TeraGrid impacted their work two of the things that users mentioned frequently were data and file transfer capabilities and access to data storage. This agreement between users and TeraGrid personnel was unusual, though. In general, when users described the benefits of TeraGrid, they did not emphasize the technological developments and common processes that TeraGrid personnel mentioned. There are several possible reasons for this. First, TeraGrid participants acknowledged that these technologies and processes were not yet working optimally. The quote below is representative of comments made in this regard.

> I think some aspects of the common software environment, a common environment where allocations get done, a common environment where user accounts get created, can all be counted as successes. I certainly think we could do a better job at all those things, but I do think that it does and has made a difference for the users. The network is a technical success story. Like everything else I can attach a caveat to it.

Other personnel who identified these types of successes were careful to note, as the interviewee quoted above did, that the solutions were not perfect. A statement made in regard to the development of parallel file systems reflects attitudes about other technologies, too: "It isn't a completely solved problem by any stretch, but we made really good progress on it." Based on other findings from this study, the caveats could reflect several things.

- technological immaturity that will improve with further development
- the challenges of system heterogeneity
- site-specific unwillingness or legitimate reasons for lack of participation, or
- a combination of one or more of these reasons

For instance, as shown in results from the user survey, user interviews, and planning workshops, users are frustrated by the difficulty of obtaining account balances and

information on the software available on individual TeraGrid resources. A second reason for different perceptions between users and TeraGrid personnel regarding these successes is that some require changes to user behavior. Users are reluctant to alter the way they carry out particular tasks unless they perceive a significant advantage or the technology makes it easy to try to new approaches. The use of MyProxy to reduce the "headaches," as one person described it, of managing private keys and certificates is an example of this. The option to login into MyProxy through the user portal without having to go through other steps increased the use of MyProxy.

> I think integrating with the user portal has been a big success for MyProxy. MyProxy has always sort of been there as a TeraGrid service, but the user portal is the first case where it's really coming into the mainstream for users; they're getting this username and password in a packet that they can use to use MyProxy, and it's hopefully starting to become a standard part of what they're doing.

But some users, particularly those with smaller allocations, are not interested in capabilities that make it possible for them to compute at different sites or to move data from one place to another other than from an RP site to their local computer. Many users are interested in computing and storing data at one or sometimes two sites. For these individuals, common processes or technological solutions that make it easier to work at multiple sites actually interfere with their activities. For example, some users mentioned the confusion caused by the fact that they received login information for all nine RP sites when they obtained a TeraGrid allocation even though they did not have accounts on all the resources. A few users also expressed annoyance over email messages they perceived as irrelevant to their use of TeraGrid such as downtime for resources on which they did not have an account. Many TeraGrid personnel were aware of these issues, and the project has taken steps to address some of them. Other issues have proved more challenging for the many reasons we have already discussed in this report.

While there was wide-spread agreement among TeraGrid personnel concerning technological and common process achievements, there were also individual TeraGrid personnel who saw particular developments less favorably than others. For example, GPFS-WAN is not mounted on all TeraGrid platforms. As in other cases, resource heterogeneity played a role in this difference between sites. A user support person said simply that GPFS and other mounted file systems sometimes caused problems for users and for the sites. Another person was more critical, however, and stated that GPFS-WAN was unstable, implying that the system that she knew others perceived as a success was not all it was made out to be. Someone else who viewed GPFS-WAN as a success believed that this view was not supported by everyone in TeraGrid because of a "not invented here" mentality.

The "not invented here" mindset provides insight into another aspect of the technologies and processes developed by TeraGrid. This is that the implementation of common processes and technological solutions were often not the result of extensive multi-site collaboration. For instance, interviewees noted that:

- TACC led development of the user portal;
- account management was based primarily on efforts by personnel at NCSA;
- GPFS-WAN was pushed forward by SDSC; and
- the centralized allocation process was built on one that existed prior to TeraGrid.

On one hand, division of labor and building on existing solutions or efforts can be efficient and effective ways to accomplish tasks, especially in a five-year project like TeraGrid where quick progress is expected and built-from-scratch solutions, in most cases, would take longer to achieve. At a basic level, the choice is to implement existing processes or solutions under development at one site across diverse institutions and architectures, which, as many of the findings in this report show is difficult, or to conduct design and development activities from scratch. The first option makes it hard to establish homogeneity across the sites because variations in environments are not taken fully into account at the start; this leads to heterogeneity, which complicates the ability of users to work at more than one site. Open-ended responses from the user survey showed that heterogeneity is a barrier to TeraGrid use. However, given the unplanned growth and dynamic nature of TeraGrid it is unlikely that all resource and policy differences could ever be accounted for completely. Retrofitting activities seem to be an inherent aspect of cyberinfrastructure. On the other hand, it is possible that starting from scratch would give participants more of a sense of shared accomplishment. We saw evidence that some Working Groups, which include at least one participant from each site, achieved camaraderie in the process of meeting their charter.

Broadening the User Base

The science gateways program was the third most frequently mentioned success by TeraGrid personnel. Interviewees viewed gateways as an effective strategy to open up TeraGrid resources to new communities of users by reducing the barriers to use.

> An engineer that's worried about nanotechnology doesn't really want to worry about which language is being used; what file system specifics; how you parallelize, and what have you. They want answers; they're engineers. To me, that's one of the biggest goals of the TeraGrid.

As with other areas, TeraGrid participants acknowledged that there were many challenges to overcome in order for gateways to achieve their full potential. For one, gateways were described as being in their infancy in the sense that there is not yet a clear or shared understanding of the gateway concept. Gateways currently provide a variety of capabilities such as access to a community specific set of applications, workflows, visualization, resource discovery, and job execution services (Wilkins-Diehr, 2006). Second, gateway developers and TeraGrid often share different concerns. TeraGrid needs to track usage, maintain security, and show scientific impact. The latter was a particular challenge since TeraGrid personnel believe they must demonstrate that use of their resources leads to "big science successes." They noted, "It's hard to show a big science success story with a bunch of little users, especially a gateway." We found that gateways are not unconcerned with issues of usage, security, and impact. However, since most are in development, their focus is on different goals such as understanding user needs and growing their user base.

In spite of the challenges, several TeraGrid personnel we interviewed echoed the sentiments of a colleague who said:

> I still think that science gateways should be what TeraGrid leads with—that and a handful of these Karniadakis sorts of things where they're really using grid technologies for full scale application runs and things like that.[27] Those are the sorts of successes that we should be driving towards. But if we want to have broad impact, the science gateways are going to be the way to do that.

The same individual quoted above also noted that once gateways are in full production they will provide a positive story about the usage of grid services since almost all of them use a grid service type of approach, including certificate-based authentication of grid services and Globus GRAM for job submission. He also cautioned, though, that TeraGrid needed to be careful not to take more credit than they deserved for the concept of science gateways.

> The gateways are mostly ongoing projects at various places. We're just providing some back-end resources. Let's be honest about what we're doing there. And there's still a good story to tell behind that—if we don't oversell the story. If we can show small growth early on and good adoption over time that is a much better story than some giant step function. It worries me that we're grabbing too much, and it's not ours to grab.

This precautionary message seemed well justified. Those who are developing science gateways do a lot of the work required to make them a viable concept. The gateways are responsible for interacting with and assessing the needs of their particular user communities and for developing some of the technologies, alone or in cooperation with TeraGrid (see also Zimmerman & Finholt, 2007). Some of the gateway developers we interviewed did feel that TeraGrid "grabs too much."

Ultimate Success

While the successes described by TeraGrid personnel are significant, they do not yet match their views of what it will mean for TeraGrid to truly succeed. Ultimate success was described by interviewees as enabling research that could not have been done elsewhere and being able to demonstrate the value of that research. In addition, TeraGrid participants believed that these outcomes would be the result of an environment that appropriately balances user needs in the present with gentle steering of users toward new ways of doing things. The sense of disappointment regarding project successes to date appeared to be the result of an inadequate link between these outcomes and the vision that most of the TeraGrid personnel we interviewed had for the project. This situation was made more difficult by the fact that few users demanded what TeraGrid was created to provide.

[27] The work of George Karniadakis, Professor of Applied Mathematics at Brown University, and his group is a grid computing success story for TeraGrid in that the Karniadkis group is able to launch jobs and have them run at multiple sites concurrently. Information about the group's research is available at http://www.cfm.brown.edu/crunch/index2.shtml.

6.3.4 Mission

Unlike the situation regarding vision, the proposal for the terascale extension articulated a mission that has been consistently and widely stated since the project was funded. TeraGrid's mission is to support science through three integrated initiatives (Catlett et al., 2008).

- **Deep: Enable Terascale/Petascale Science**: TeraGrid will enable scientists to pursue scientific discovery through an integrated set of Terascale resources and services
- **Wide: Empower Communities**: TeraGrid will make Terascale resources and services broadly available through partnerships with community-driven service providers
- **Open: Provide an Extensible Foundation for Cyberinfrastructure**: TeraGrid will provide, and use where provided by others, a set of foundational services and resources to support nation-wide cyberinfrastructure, using open standards, policy, and processes.

Personnel were able to articulate the mission as stated above, which demonstrates that it had been clearly communicated. Further, interviewees did not appear to disagree with the mission. Instead, the concern was over how to prioritize the three facets of the mission, particularly, the first two, since resources are limited. One person succinctly summed up the view of many personnel when he said, "The TeraGrid vision of deep, wide, and open boils down to, 'Let's make everybody happy.'" The many issues that we have already discussed in this report intensified the dilemma caused by conflicting demands stemming from TeraGrid's three-pronged mission.

> We've only got so many resources. We have to decide and prioritize. Let's pick three communities and do a good job working with them and let the rest of the people, the masses, just benefit from the results of what we do there. We can't help everybody at the same time.

We analyze user needs in detail in a later part of this report. For now, suffice it to say that trying to meet the requirements of diverse types of users and enable multiple modes of usage often places conflicting demands on existing technologies and policies. The challenge is not so much that TeraGrid personnel do not understand what users need and want; the difficulty is in knowing who to pay attention to, when, how, and to what degree (see also section 8.1.1).

> We've got the traditional users, and if you look at the way the allocations break down, a good third of the total allocated usage boil downs to three or four users. You don't want to do anything to mess those guys up. Those are the bread and butter guys. Regardless of output, regardless of… We might want to look at it regardless of the vision for gateways or anything else—there are a small handful of scientific users who drive and populate the TeraGrid. And we can talk about deep, wide, and open all we want to. The reality is deep is the only one that really matters at this stage. Everything else is still something that we want to do. Deep is where it's been for 20 years, and it's where it still is today.

As the quote above shows, large users are important to the resource providers, and they exert a strong influence. We discerned two related reasons for this. First, a metric of success for

HPC sites is utilization of resources. Since supercomputers are costly and unique resources, it is important to demonstrate that they are in use as close to 100% of the time as possible. Thus, policies are designed to guarantee maximization of use. This makes it hard to meet needs for on-demand computing, which is often required for weather prediction and emergency response, and to serve new classes of users such as some experimentalists who desire quick results in order to test hypotheses and make decisions on experimental conditions.

> We have two kinds of modes. One is people complaining about wait time, and the other is idle processors. Trying to find that sweet spot. The way everybody wants to run a computer is on-demand. "I need 128 processors now! Let's go." The problem is everybody can't be the most important and have preemptive access. ... The only way to really give people the best quality of service is to really under-allocate the machine. And we do have control over that. So, it would be like, "The machine's idle most of the time, but, boy, the wait times are great." We can't afford to do that because we have $10 million in this machine, and we want to get the most utilization out of it as possible.

In addition, it is important to show that use takes advantage of the architecture of a particular resource and that codes run efficiently. Utilizing large amounts of time on a specialized resource in an efficient manner is not easy. Most of the projects with large allocations that we studied employ multiple people and direct significant time and effort toward computing. Even though the findings from the user workshop (Zimmerman & Finholt, 2006), the *TeraGrid User Survey*, and interviews with users show that users with large allocations have needs for education, training, and support due to the turnover of graduate students and postdocs, it is the case that supporting the needs of a small number of large projects is more feasible than serving thousands of “little” users.

> There's lots of users doing little stuff, and then there are these "grand challenge" people who have a lot of clout and are very visible. And we've always kind of paid special attention to them. And we've always had to weigh and balance how much effort do you give single users and their project versus these thousands of users out here. And you can't do it all.

Even though education and training need exist in projects with large allocations, the fact that these projects usually include a team of people means that there is support and training available within the group. The number of users needing education and support could explode as usage of science gateways increases. Even if gateways provide these services for their users, TeraGrid personnel will have to, at minimum, troubleshoot problems that involve the use of their resources and provide some of the education and training for new and often less experienced communities of users.

Although interviewees described the project's most significant obstacles as being social and organizational in nature, the technical complexities of tying together distributed, heterogeneous systems to serve the needs of multiple users are significant.

> We're adding grid interfaces to systems that are already well defined; we already have processes in place, and so building the glue for that in terms of both policies and mechanisms has been a challenge throughout the project.

As we noted previously, many personnel were not convinced that there was sufficient demand for grid computing to justify the effort and time spent on this task. The growth in the number of RPs and the diversity in site hardware and policies added to the technical difficulty of achieving this goal.

> A good example is that initially many things were designed—I'll use the account management process as an example—assuming there are four sites, and that's all there would ever be. And there were certain assumptions associated with that, which all went away a year later once ETF was born. So, a bunch of things were changed to add that functionality in there, and then it changed again. The program keeps changing, and the way things were designed—initially under assumptions that no longer held—were broken. And they were done fairly quickly at that point just to try to get them done to the point where we could deploy and be successful and provide the resources. And it's never been possible to sort of go back and fix it right.

Science gateways further increased the technical complexity. Besides the technical challenges they created, existing policies and processes, as described in the quote below, were also a poor fit for the needs of science gateways.

> They weren't even on the radar when we made the transition into production. So, that's a new development effort, but it necessitates going back and re-architecting lots of things that we made during the construction phase that we now need to go back and deal with.

We have analyzed tensions in priorities between two of the three elements of TeraGrid's mission, but we have not discussed the *open* aspect of the mission. This is due to the fact that it was rarely mentioned in interviews or was it a topic at meetings we observed. For this reason, given the limited time of our study, we chose to focus on issues and topics that interviewees identified as most important.

6.4 Research versus Production Infrastructure

The last tension we analyze is the one between the desire for infrastructural reliability and stability that makes it easier to accomplish work, particularly on the part of users, and the research and development that are often necessary to create a distributed cyberinfrastructure. TeraGrid personnel expressed three opinions on this issue: 1) TeraGrid is foremost a research and development project; 2) TeraGrid should be a stable and reliable cyberinfrastructure; and 3) TeraGrid is both research and development and cyberinfrastructure. As an interviewee told one of us:

> If I were in your role, I would really try to focus on what do I think the barriers are to users using TeraGrid and the tension between computer science research, deploying

> computer infrastructure, and then actually getting the science done, and what TeraGrid's balance of these three things is supposed to be. … In my mind, different people have different views of those three things.

No one we interviewed advocated for research at the expense of users' abilities to conduct research. As the quote above states clearly, there were different views, however, about how far the balance could swing between R & D and the provision of a production environment. These views are captured below in the words of different interviewees.

> TeraGrid is not a research project. It's infrastructure. Infrastructure is supposed to work, and in many ways it doesn't work.

> This is not a development project; this is a research project. And that means there's risk. It means it can fail. In a sense of, you know, strictly deliverables.

> Given the ideal environment, I think a good balance of the two is workable.

Studies of other cyberinfrastructure projects have also noted the tension between research and development and the provision of a robust and reliable infrastructure (Lawrence, 2006; Ribes & Bowker, 2008; Ribes & Finholt, 2007).

6.5 Summary: The TeraGrid Collaboration

It is clear from our interviews with TeraGrid personnel and from our observations that many in TeraGrid

- are aware of the tensions that exist;
- share frustration over the lack of clarity regarding the project's vision and priorities;
- have substantial knowledge of and dedication to their users; and
- believe that TeraGrid could (and has), enable new kinds of science, technological advances, and be a model for distributed delivery and support of cyberinfrastructure.

Given this, why does the knowledge that exists in the project often not get shared, discussed, or acted upon? In his book on strategic planning, John Bryson (2005) stated that it is important to obtain an answer to this question. We identified three factors that impeded the sharing of knowledge within the TeraGrid virtual organization. We have touched upon these in other parts of this section, but we bring them together here in answer to the question above.

One set of explanatory factors stems from challenges that the growing body of literature shows is common to many virtual organizations. Of these, trust can be among the most difficult to achieve, and a lack of trust can make it harder to arrive at a shared vision and to balance autonomy and interdependence and collaboration and competition.

> Trust is a big issue. ... And nobody wants to talk about it because everyone wants to appear congenial, and everybody needs to appear as if they're team players. ... The NSF does a disservice to its goals by using the competition stick as the only tool that

> they use to try to shake out who should get what. What happens is the PIs go into battling for dollars. They're battling for survival; they're battling to keep staff, and it really has a big impact on how well partnerships can be built and trust can be built.

One of the project participants noted that often trust is "based on the experiences of actually working with people." TeraGrid personnel mentioned that one of the project's main successes was the spirit of collaboration that it engendered among colleagues at other sites. This was a pleasant surprise to many of the individuals that noted this success, and it may bode well for TeraGrid's future. As new sites become part of the project, though, trust must be continually developed.

A second category of explanations arises from pressures to present a united front, at least to the outside world, and to show results quickly and constantly. One affect of these pressures is a disinclination to take risks.

> I definitely see that from a management perspective that sometimes you can take a risk and you might get a lot of bang for your buck, but there's disincentive to do that and it's just a little bit easier and more comfortable to just try to stick to the standard party line and move forward—maybe incrementally—that way. Where, if you took a flyer on a couple of other things, you could take some bigger leaps forward. ... I think they have been successful but, again, I think, going forward that it would be nice to figure out where we can take some risks as a group and not have finger pointing.

The need to demonstrate results exacerbated the tension between those who perceived TeraGrid as primarily a research and development project and those who viewed its main purpose to provide stable and reliable cyberinfrastructure. Further, it is a strain on participants to present a united front in the face of ongoing competition and diverse organizational cultures, missions, philosophies, and capabilities.

These challenges are not unique to TeraGrid; accountability and public scrutiny are inherent to large-scale, multi-million dollar projects (e.g., Collins, 2003). In addition, the funding that RP sites receive for TeraGrid activities is only one reason for their participation in this endeavor. TeraGrid is a significant part of NSF's cyberinfrastructure program and one that organizations either want to be involved in or cannot afford to be left out of. In regard to the former, a person at a site that was not one of the original NSF-funded centers described his organization's reason for wanting to participate in TeraGrid:

> It starts tying us into NSF in a big way and cyberinfrastructure, which is, of course, a big emphasis right now. And it also aligned us with a lot of places around the country that are also doing a lot of "good things" in cyberinfrastrucure. So, I saw it as a way of us really moving quickly toward our goal, and it has really helped us do that.

The diversity of the organizations that comprise the TeraGrid is a source of innovation, but it also raises questions about the roles and responsibilities of sites and what it means to be a member of TeraGrid

Third, there is a notion that the leadership, the management structure, or both have been unable to overcome the many sources of tension. The result is that people are cautious and often hesitant to speak out.

> So, you are constantly playing that game: competing today, but tomorrow you're on a conference call on the TeraGrid stuff, and you have to collaborate. I think that the management can deal with it better. The staff don't know what they can say because they don't know what the negotiations are and what the thinking is because it's changed anyway. They just don't know how they're supposed to interact, so they tend to clam up because they want to try to protect themselves and their own site. ... But that means that the conversations that really need to happen in the TeraGrid don't happen, and the people aren't as open as they need to be and for good reason.

The findings from our study of TeraGrid show how difficult it is to juggle the tensions and competing demands inherent in a virtual organization even when there is widespread agreement about what is ultimately the most important goal—in this case, enabling the work of scientists and engineers. In the midst of these difficulties, and the stresses they create for the participants, positive outcomes can be achieved. Some of the most significant successes may be ones that were unexpected. The ability to collaborate across sites, to learn from people at other institutions to improve and enhance local procedures, and to work together to achieve a challenging technical goal within an environment of competition and lack of trust should be viewed as important successes that may last beyond this particular project.

7. Grid Computing

In this section we analyze reasons for the slow adoption of grid computing by current TeraGrid users. Results from this part of the study provide a useful bridge between our analysis of the TeraGrid collaboration and the next section, which synthesizes data we collected on user needs. We describe technical and practical reasons for the limited interest in grid computing. We find that the concept of grid computing is evolving to encompass a broader range of distributed computing tasks. This is a positive development that can be attributed partially to technologies developed by TeraGrid and science gateways.

7.1 Defining Grid Computing

What is grid computing? Although interviewees told us that there was not a consensus on the definition, there was more agreement than it appeared.[28] We found that the debate was not so much over the definition of grid computing as it was about whether it was worth the human resources and funds that were being expended to achieve it, particularly since only a small number of users were interested or able to compute in this way. One of the CI experts we interviewed described what we found to be the common conception of grid computing at the time TeraGrid was initiated.

[28] There is a significant amount of literature on grid computing (e.g., Berman, Fox, & Hey, 2003; Foster & Kessleman, 2004). We do not discuss definitions of grid computing found in these resources. We are concerned here only with the views of those interviewed as part of this study.

> The idea was that you would have load balancing at a very high level among national supercomputing centers and to enable researchers to run jobs across multiple systems at the same time.

The ability to run at more than one site concurrently was referred to by most interviewees as *co-scheduling*. Further, the prevailing definition of the time suggested that the number of jobs to be run was large and required a significant amount of computing power; controlled execution was employed to send jobs to many places in order to achieve load balancing.

Interviewees, particularly TeraGrid personnel and some gateway developers, advocated for *meta-scheduling* rather than *co-scheduling*. *Meta-scheduling* was defined as the ability to submit a job and have it run on the first available resource that is appropriate. The fact that people distinguished between meta- and co-scheduling is further evidence that there was a common understanding of grid computing. Again, the debate was primarily over whether co-scheduling should be the focus. The quote below, by a TeraGrid participant, is representative of what others said.

> There's been a lot of effort trying to do what they call meta-scheduling, but what they really mean is co-scheduling. You're running at 2, 3 or 4 sites concurrently. That's a wonderful goal for the 5 people who care about it. A lot more people care about submitting a job and getting their job back, and they don't care where it runs. This is really meta-scheduling, where I can submit a job and it might run here, or it might run, there, or it might run there. It runs where it can run first. A lot more people care about that. It's a lot less sexy, but it gets science done, which is what researchers care about.

A number of TeraGrid personnel voiced frustration over the fact that the project had not yet made it possible to do meta-scheduling. They felt that it was technically achievable and that there had been adequate time for this goal to be reached.

7.2 Barriers to Grid Computing

Interviewees in all categories noted that grid computing as defined above (i.e., co-scheduling) is hard to achieve for a number of technical and practical reasons. These are in addition to ones we have already mentioned such as the heterogeneity of the resources and the fact that the many traditional users of national centers are not interested in grid computing; they are interested in exploiting the unique aspect of particular resources and have optimized their codes to run efficiently on specific platforms. Below, we list other important factors that affect the adoption of grid computing. We do not elaborate on the technical barriers since they can be quite complex, and they are well understood by computer scientists, users, NSF, and other HPC experts.

- Communication needs of codes

  Codes that require a lot of communication between the processors are poor candidates for grid computing because latency between sites can severely affect their performance. Even

if a particular code can be rewritten to run on the grid, it may not be worthwhile—not all jobs are appropriate for the grid.

- Costs of getting into grid computing

Some scientists and research groups have been working on one code for many years. It takes a long time to develop and modify codes so they run well on one or more architectures. Redesigning legacy codes requires a lot of effort and knowledge that can be difficult to obtain. For example, if a particular code has been written by graduate students, the PI may lack the deep knowledge of the code required to adapt it to the grid. In addition, unless users perceive a significant advantage to grid computing, they will not change their practice. One reason for this is that researchers want to control as many factors in the compute environment as possible, so they can, for instance, assess the source of unexpected results: Is it an error in the code, or the job processing, or is it an interesting new finding? Other scientists rely on third-party software, in which case they may not be able to modify the code because of intellectual property considerations.

- Middleware

Middleware was described as being unstable, primitive, and hard to use. In addition, middleware is not transparent. Users must understand all the details because many different problems can occur and identifying the point at which trouble arose can be difficult. This also makes it hard for RP sites to support grid computing. As noted previously, since most science gateways use a grid services type of approach they may help to reduce many of these barriers.

- Logistics

It requires considerable manual effort to compute concurrently at different sites. For example, a user who wants to compute at Purdue, SDSC, and TACC must interact with three different job schedulers. Since there is not an automated way to do this, the process is labor intensive and involves personal communication and hands on actions. An experienced user of grid computing, who was also a satisfied TeraGrid user, said, "I don't care how it works. You can call other people, email them, but I want to contact one person and get information about why my code is not running, why my reservation is not going well, or how to solve some particular problem that is related—not to one supercomputing center—but to several supercomputing centers." Science gateway developers expressed sentiments similar to this one. One stated, "Grid computing is primarily a problem in scientific administration." The logistical difficulties and heterogeneous procedures across sites also help to explain why some users, including gateway developers, did not perceive TeraGrid as a cohesive collaboration.

- Security

Obtaining passwords and grid certificates, renewing certificates, and managing all this information can be a barrier to grid computing.

According to interviewees, most users who run into one or more of these barriers will abandon future attempts at grid computing. Even researchers who are proponents of grid computing acknowledged many of these barriers.

> We're saying there are codes and middleware and infrastructure that exist that could enhance people's productivity, but there are a lot of users that don't even recognize that today. And it's just because of the scale of things and the complexities that have appeared in the past—some of them usability issues. There are people who've effectively burned their hands with some of the early middleware. If you're a scientist, you have objectives to meet, and you're not going to waste too much time. If you wasted a lot of time before, how are we going to persuade you to go back again this time and try it?

Still, TeraGrid has enabled some high-profile grid computing successes, and interviewees in all categories recognized that grid computing is an appropriate usage mode in some areas and for some users.

7.3 The Evolution of Grid Computing

Based on interviews conducted during the course of this study, we find that the concept of grid computing as defined in the opening of this section has broadened considerably over time. This is viewed by most as a positive development. The main changes we observed were increased emphases on *workflow*, distributed file systems, and data movement, storage, and visualization. There were also subtle changes in language. Instead of the phrase "grid computing," people spoke more of the "grid" to mean an entity that enables multiple and distributed tasks. It is difficult to identify the exact causes and timing of this conceptual shift. It is likely a combination of several factors, including

- technological advancement;
- the ways−often unanticipated−that users have employed the capabilities developed by TeraGrid;
- the influence of science gateways in helping to adapt and shape the infrastructure to accommodate the needs of their particular user communities;
- the ever growing amounts of data to move, manage, and analyze; and
- the difficulty of maintaining a homogeneous set of resources.

One of the CI experts we interviewed described the evolution of grid computing and his belief that the vision of TeraGrid could remain intact in the face of changing technology.

> Now, 6 or 7 years later technologies have changed and needs have changed, and the overarching picture that is used now is no longer the backplane. The original backplane vision was very hard to sell because of the speed of light and other issues with latency. Given the important emphasis of data cyberinfrastructure, the overarching vision is a storage area network grid file system. It’s a distributed file system, distributed on a grand scale.

Several users stated that the main benefit of having multiple RP sites connected together has been the ability to quickly move huge datasets around from one computer to another. Another CI expert explained that he perceived the original vision for TeraGrid as being about

distributed computing cycles, whereas, he said, "Now I think it's really more about storage and visualization and dynamic networks that you can throw around to support those requirements." The importance of workflow was mentioned most often by gateway developers and TeraGrid personnel. The quote below is from a TeraGrid participant.

> The thing we've been trying to lobby for is not so much the run anywhere that there's time, but sort of embracing the heterogeneity. There people are talking about workflow. I think in ultimate when this is successful, it's when somebody can sit down and construct a workflow that includes where their data is going to be, where it needs to get, what job needs to run, what data needs to come out and get back to maybe someplace else. … I think that's when success would come.

A gateway developer felt that the ability to automate workflows would attract more researchers to the TeraGrid. He noted that physicists and chemists, with their long history of HPC use, are willing and able to surmount usage barriers. Researchers in other areas, however, will not be attracted to HPC until they perceive how it will benefit their work. A workflow that automates many otherwise manual steps can make HPC easier to use and help new user communities find value in HPC. The gateways work with researchers to develop the workflows and to hide the complexity of the many steps required to execute the workflow.

7.4 Summary: Grid Computing

The original notion of grid computing as running at more than one site concurrently has evolved into the idea of employing *the grid* to enable a variety of distributed tasks and modes of usage. This is viewed as a positive development. Co-scheduling is perceived as only one way to use the grid. It is valuable to those who can take advantage of it, but the effort put into co-scheduling should be weighed against options such as meta-scheduling, which many believe would benefit a large number of users.

8. TeraGrid User Needs

> Since TeraGrid is a national facility serving the broad spectrum of science and technology, how you allow that system to meet such diverse needs is going to be an overarching challenge. And no one pretends to have the answer. It's actually an evolving field in itself. —CI expert

Historically, the use of HPC resources has been reserved for specialized scientific and engineering applications that must handle very large databases or do a great amount of computation. HPC users will increase in number and diversity as more research fields become data and computation intensive. Ecologists, for instance, have not been traditional users of HPC resources, but as their field moves toward big science with teams of researchers from multiple disciplines and sophisticated instruments that collect large amounts of data, they will need different tools and technologies to manage, store, and analyze the data they will collect (Borgman, Wallis, & Enyedy, 2007). Similar changes are occurring in other domains, including the humanities and social sciences (Hey & Trefethen, 2008; Lawrence & Zimmerman, 2007b). In fact, the compute, data, and visualization resources available through TeraGrid and other HPC providers have already begun to play a role in many disciplines as

evidenced in part by the development of science gateways (Wilkins-Diehr, 2006; Zimmerman & Finholt, 2007).

One of the goals of the TeraGrid evaluation study was to assess TeraGrid's progress in meeting user needs. To reiterate, the questions that drove this part of the investigation were:

- What factors affect users' computing needs and requirements?
- How are the needs of users expected to change over the next five years?
- What factors affect users' behavior as it relates to their use or non-use of computing resources and services?

Users included individuals who *currently use* TeraGrid and *target users*, researchers who do not yet use TeraGrid, but who are likely to do so in the future. Since it is difficult to study users who, in large part, do not yet exist, we interviewed developers of science gateways to better understand the needs of potential future users. The timing of the study enabled us to get a snapshot of the first generation of HPC users, more recent users, and emerging user communities.

We employed methods from the field of user-centered design (UCD) to analyze user needs. User-centered design emphasizes the importance of understanding the work practices of users and maintaining ongoing interaction with them as part of an iterative process of analysis, design, development, and system implementation (Norman & Draper, 1986; Olson, Finholt & Teasley, 2000). Given the scope of TeraGrid and its user base and the time limitations of the study, a full UCD process was beyond the resources of this investigation. Therefore, we focused on an analysis of user needs—the first step in UCD. The result of this analysis is an understanding of the factors that influence users' needs, affect their decisions about how or whether to use TeraGrid, and help to explain variations in their needs. We also identified issues that are likely to affect user needs in the future. As we have noted elsewhere, differences in the requirements of users can be a source of tension that makes it hard to set priorities. While these issues are difficult to fully resolve, a clearer understanding of user needs should aid NSF, TeraGrid, and other stakeholders to develop strategies to serve users who address many types of research problems and have varying levels of HPC experience, knowledge, and skill.

Interestingly, it also became clear to us as our investigation progressed that a substantial body of information exists about user needs. While this is primarily the case for fields that, to some degree, already employ HPC resources, there is also a growing amount of information available on the needs of communities that have not traditionally employed HPC. This information appears in numerous reports, many of which present findings from workshops sponsored by various NSF divisions and programs. The workshops were structured to define user requirements for cyberinfrastructure, including HPC, in terms of research drivers that determine needs for hardware; network infrastructure; data; human resource capacity and

development; and algorithms, models, and software. The examples below illustrate the range of areas covered by the workshops.[29]

- *Cyberinfrastructure for Environmental Research and Education*, 2002
- *Cyberinfrastructure for the Atmospheric Sciences in the 21st Century*, 2004
- *Final Report: NSF SCE-CISE Workshop on Cyberinfrastructure and the Social Sciences*, 2005
- *Identifying Major Scientific Challenges in the Mathematical and Physical Sciences and their Cyberinfrastructure Needs*, 2004
- *Materials Research Cyberscience Enabled by Cyberinfrastructure*, 2004
- *Our Cultural Commonwealth: The Report of the American Council of Learned Societies Commission on Cyberinfrastructure for the Humanities and Social Sciences*, 2006[30]
- *Summary Report: NSF EPSCoR Cyberinfrastructure Workshop*, 2006

Elsewhere in this section, we refer to the reports described above as *cyberinfrastructure reports.* Another excellent and comprehensive source of information that we drew from is the volume edited by Graham, Snir, and Patterson (2005), which presents findings from a National Research Council committee's assessment of U.S. supercomputing capabilities.

Recently, there has been a focus on examining the opportunities for research progress that could be enabled by petascale computational capability and determining the steps to prepare for the use of these resources when they become available (e.g., Snavely, Jacobs, & Bader, 2006; Yeung et al., 2007).

The plethora of existing information raised new questions for our study as it related to user needs.

- Is it the case that existing information is not well known or adequately organized and, therefore, not used? Or, is use of the knowledge limited by other factors?
- Would more effective use of this information provide better guidance for TeraGrid hardware procurement, software development, user support, and education, training, and outreach programs?
- Does the available information address all the factors that influence user needs and behavior? If not, what is missing?

In this section, we address the questions above along with those we formulated at the start of the investigation. We begin by drawing from and expanding on results presented in section 6 of the report − The TeraGrid Collaboration. We find that unresolved tensions over how to balance the needs of different types of users, the lack of a clear vision, and the broad definition of a TeraGrid user leads to several dilemmas that hinder the use of available information regarding user needs. Second, we present the results of the user needs analysis

[29] A more complete list of reports appears as Appendix B in NSF's 2007 publication, *Cyberinfrastructure Vision for 21st Century Discovery*.
[30] The Commission was supported by The Andrew W. Mellon Foundation.

we conducted. Our results confirm and support those from prior activities. In addition, our investigation identified other important aspects of user needs. We conclude the section with an analysis of areas that will shape user needs in the future.

8.1 Dilemmas

In spite of the pressure for TeraGrid to serve a larger and more diverse user base, it is difficult for TeraGrid to meet these demands. The literature on the historical, social, and technical aspects of infrastructure has identified several reasons for this that are relevant to TeraGrid (e.g., Edwards et al., 2007; Star, 1999). These include mismatches between potential new users and the TeraGrid infrastructure, the inability of TeraGrid to interact with many individual users, and the conservative influence of the "installed base" of users, technologies, and institutions, which pushes against change (Monteiro, 1998, p. 229). The tension between growth and expansion on one hand and existing technical and non-technical elements on the other, create three dilemmas for TeraGrid.

8.1.1 Dilemma #1: Resources must be used, but support is limited

A large percentage of TeraGrid allocations are awarded to a small fraction of principal investigators. For example, according to TeraGrid quarterly statistics for January-March 2007, the top twenty PIs as measured by usage consumed almost two-thirds of the normalized units (NUs) utilized by all 879 PIs with allocations during this time period. A similar situation applies to domains. Figure 2 shows that projects in six broad areas accounted for more than 98% of TeraGrid usage as measured by multiple factors. In addition, the PIs associated with the top twenty projects by usage during this quarter were from the seven broad disciplinary categories listed in figure 2 (e.g., atmospheric sciences, chemistry), plus a user in ocean sciences. Given these statistics, a strategy focused on the needs of individuals associated with projects and disciplinary areas with the greatest TeraGrid usage is compelling. Of course, TeraGrid's stated mission is broader than this; NSF has encouraged TeraGrid to increase participation in terms of disciplines, usage modes, and gender and race; and new communities of users with different types of needs are putting pressure on RP policies and technologies. In spite of these demands, the statistics above help to explain the dilemma TeraGrid faces in trying to broaden and increase its user base.

First, on the TeraGrid side, utilization of resources is a measure of RP success. Second, on the user side, it requires people, time, and expertise to utilize millions of service units. Therefore, it is not surprising that those who can consume significant numbers of cycles are considered the "bread and better guys" (see section 6.3.4).[31] Further, as a TeraGrid user consultant described, these users support many of their own needs.[32]

> The big users typically are very sophisticated and don't have that many questions, and when they do, it's not related to their code it's related to our quality of environment or

[31] To say "guys" is not totally accurate: the top twenty users in the first quarter of 2007 included one woman. Some of the users associated with the top twenty projects are probably female as well. TeraGrid does not keep statistics on the gender of users, but other information, including results from the user survey indicate that the majority of TeraGrid users are male.

[32] There are exceptions to this, which we discuss later in this section.

> something like that. Because a lot of them, like the MILC group and the NAMD group, have computer scientists who are working with them to make their code run faster. So, they spend a lot of time finely tuning their code to run on various different machines, and they're very knowledgeable. … Most of the problems they have are more operational problems where it's something that is a problem on the system, something is not working right, or something is not configured right.[33]

Results from the *TeraGrid User Survey* showed that there is a relationship between frequency of TeraGrid use, which is related to allocation size, and user satisfaction. Individuals affiliated with groups with large allocations have needs for, and in many cases would like, more access to education and training and support, but there is generally an option to obtain help within the project.

Third, more users place strains on TeraGrid personnel, policies, and procedures, particularly if many of these additional users have less experience in using HPC resources, and therefore require a higher level of support. Technical complexity also increases when a wider range of hardware and software are needed to meet user demands. These issues were summarized by the PI of a science gateway project, who was critical of TeraGrid's efforts to meet the needs of a broader base of users.

> The new communities that would benefit from high-performance computing find a real impedance mismatch between the way the machines are fielded, the way the staff are there to help them understand how you would use this in your science project, and also the way the machines are actually architected—the amount of memory, the portability of codes to that particular architecture.

TeraGrid personnel recognized these challenges (see section 6.3.4), but they are hard to resolve. Because so many resources at RP sites are considered TeraGrid resources and because there is a TeraGrid process by which time is allocated on these resources, the needs of large users are a significant concern for TeraGrid. Even if these users were not considered TeraGrid users by definition, meeting their needs would still be paramount for the reasons we have identified.

[33] The MILC (MIMD Lattice Computation) Collaboration is one of the largest users of open-science computing in the world. By April 2008, they had already consumed 19 million computing hours at TACC alone (Dubrow, 2008). NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems. It was developed by the group of physicist Klaus Schulten at the Beckman Institute for Advanced Science and Technology. Schulten's group is also one of the largest users of TeraGrid resources. For the first quarter of 2007, PIs Robert Sugar of the MILC Collaboration and Klaus Schulten were the second and third largest users, respectively, of TeraGrid as measured by usage.

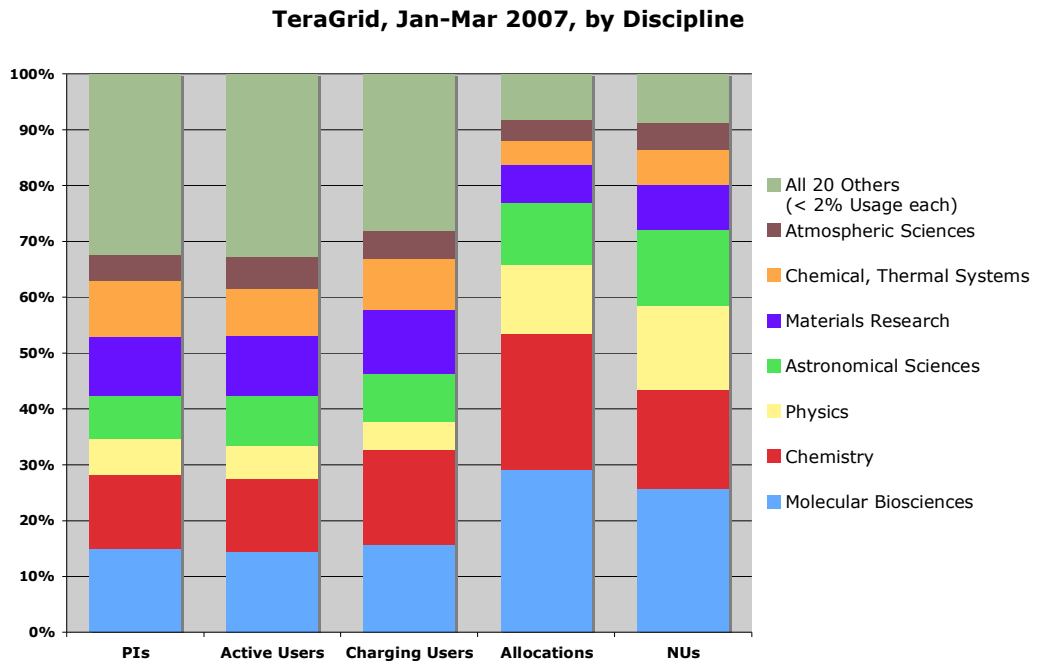

Figure 2: TeraGrid Usage by Discipline (figure created by David Hart, SDSC)

8.1.2 Dilemma #2: Utilization versus Impact

In spite of the fact that almost all usage of TeraGrid is accounted for by a small number of PIs, the remaining time supports numerous scientists in a variety of areas. Figure 2 shows that the 20 disciplines that comprise less than 2% usage in terms of number of PIs, active users, and charging users are equal to roughly 30% of the total in each of these areas even though they make up less than 10% of the allocated time and NUs utilized during this quarter. Although they are small in terms of resources consumed, taken together, these researchers' contributions to knowledge production and scientific advancement may equal, or even surpass the output of all projects with substantially larger allocations. The scientific benefit derived from use of TeraGrid is an extremely difficult outcome to assess, however. The report from the TeraGrid Impact RAT (2006) suggested ways to evaluate the merit of projects that utilize TeraGrid resources, but it remains an intractable problem.[34] It is even harder to evaluate the relative impact of different types of science. For example, how does one compare the results of research that contributes to basic theory with applied research that may have more immediate economic benefit? Or, are the many small jobs that could be enabled by meta-computing equal in research output to jobs run concurrently at more than on RP site? Because such questions are nearly impossible to answer, TeraGrid personnel and

[34] TeraGrid LRAC and MRAC allocation committees are instructed not to judge the scientific merit of the work described in an allocation proposal as this is the role of the peer review process under which the research was funded. The task of allocation committees is to evaluate the quality of the past and proposed mapping of scientific progress to the TeraGrid resources. Yet, TeraGrid and users are expected to provide evidence of the scientific results produced through the use of TeraGrid resources.

some users stated that science success is often equated with size such that the larger the simulation, for example, the more likely it is to be touted as a successful scientific use of resources. This situation creates tensions between users.

> We do science-based engineering to address important engineering problems using computational methods. It's different than doing pure theoretical science. Engineering problems are a lot more complicated.

The user above was referring to the difference between applied problems and basic research. He perceived a bias in queue policies toward researchers doing what he described as "one huge computation," whereas he needed to run "tens of thousands of little ones." A quote in the next section, which was taken from an interview with a theoretical high-energy physicist, is indicative of another perspective—that the use of TeraGrid should be limited to jobs that match a machine's capabilities.

8.1.3 Dilemma #3: Determining Appropriate Use

The ability to efficiently and effectively exploit the unique capabilities of HPC resources has traditionally determined what qualifies as "appropriate use" of these resources. Should this also be the metric for TeraGrid? It is difficult to answer this question without a clear vision and a robust definition of TeraGrid and a TeraGrid user.

As noted earlier, 879 PIs had TeraGrid accounts during the first quarter of 2007. These researchers use the resources in different ways. The theoretical high-energy physicist we mentioned above stated:

> When I got onto one of the TeraGrid machines, I was aghast at the number of people running single processor jobs. They could have workstations at their own university doing the same thing. So, why was this done? I think the centers are under pressure to show the resource is being used, so they keep lowering the bar, until finally the people using the resource are not using it for what it was built for.

This LRAC user expressed a sentiment that we heard from other large users and from TeraGrid personnel; similar comments were also made in response to open-ended questions on the user survey (see Appendix D, part 2 of this report). The issue is more complex than the statement above implies, however. While no one argued for "inappropriate" use of resources, interviewees recognized that it is not a simple matter to determine what is appropriate. For one, as the physicist stated, it is important that a multi-million dollar program is perceived as serving more than a small percentage of the research community. Second, many users run more than one type of job, and these jobs can have different computational requirements. It is not simply that large users run "big" jobs and small users run "small" ones. For example, the engineer quoted above, who had an LRAC award, solves hundreds or thousands of simulations, varying one parameter or another to discern which are most sensitive, so he can then make comparisons with experimental data. Third, a comment we heard consistently from the users we interviewed is the value of having computational resources that are

available (i.e., *open*) to researchers or educators at U.S. academic or nonprofit institutions whether or not they receive funding from NSF.[35] This was noted even by users who were dissatisfied with TeraGrid. Results from the *TeraGrid User Survey* showed that only half of the respondents have HPC resources available at their institutions and 14% have access to state or regional resources. Interviews indicate that even when these options exist, the quality in terms of resource stability and reliability and user support are variable with some local centers being very strong and others less so. Departmental or project resources are preferred by many users, but some funding agencies do not support such purchases, and so users must rely on NSF-funded resources. As one user noted:

> When we get funded by the NIH to do a lot of this work, they actually assume we can do the calculations we do without giving us money for it. So, having these NSF centers is very important because they allow us to do these calculations.

Departmental and project resources also incur expenses for maintenance and support, space, and power and cooling costs. Energy demands, in particular, are becoming a serious issue for institutions (Hacker & Wheeler, 2007). In addition, users noted that access to NSF-funded resources by foreign nationals is a major benefit since it is common in research environments to have students, postdocs, and faculty who are not U.S. citizens. Interviewees who mentioned this generally believed that DOE facilities, for example, are not open to foreign nationals.[36] Some interviewees who had allocations at a DOE center also noted that stringent security regulations can be a barrier to use. For all these reasons, it is difficult for NSF-supported RP sites to strictly limit resource use to particular types of jobs.

8.1.4 Addressing the Dilemmas

Until these dilemmas are resolved, existing information on user needs, including those based on data collected in this study and presented in the next section, will not be utilized effectively. This statement is not intended to downplay the complexity of meeting the needs of diverse users at a time when the demand for HPC is growing, or to ignore the urgency for improved methods to organize the substantial information that exists and to find ways to employ it strategically. In addition, the effective use of user needs data will rely on ongoing assessment of user needs and on interactions between multiple parities, including NSF, TeraGrid, users, and other stakeholders.

---

[35] Details regarding eligibility to use TeraGrid resources are available at http://www.teragrid.org/userinfo/access/allocations.php. DAC allocations are available to K-12 teachers for classroom use, and international researchers can use TeraGrid resources if a U.S. researcher is the PI on the allocation award.

[36] DOE Leadership Computing Facilities are managed as open national resources. While they are available to all researchers, including foreign nationals, the focus of their use is on "computationally intensive, high-impact scientific applications" (U.S. Department of Energy, Office of Science, 2007, n.p.) In addition, DOE's INCITE (Innovative and Novel Computational Impact on Theory and Experiment) program reserves ten percent of the DOE Office of Science high-end computing resources to allocate to the broad scientific community, including private industry, with no requirement of DOE funding.

8.2 User Needs Analysis

The goal of the user needs analysis was to gain an integrated picture of users and their cultural, organizational, and research contexts (Lindgaard et al., 2006). Specifically, we collected data from individuals representative of current and target TeraGrid user populations in order to understand the needs of different types of users and factors that influence user behavior. The results from the user survey offer a generalizable view of current TeraGrid users while the interview data provide in-depth information and help explain some of the survey findings. Science gateway developers are also users of TeraGrid, but since their needs are different from those of individual users we discuss them in a later section. Interviews with science gateway developers did, however, provide us with information about the likely needs of future TeraGrid users as did interviews with new TeraGrid users.

We approached research questions related to the needs of TeraGrid users from several perspectives. First, we asked questions intended to uncover areas where users currently spend time that does not come under the heading of "doing science." These are activities and tasks that are candidates for automation or other improvement. Second, we queried users about future scientific problems they would like to address or specific questions they want to answer and the challenges to doing so whether or not they were related to computational resources or services. This helped us to identify needs that are within the traditional realm of HPC providers to meet; ones that TeraGrid does not currently play a role in fulfilling, but might consider doing so, alone or in collaboration with others; and needs that should be addressed by others such as users' local institutions. Third, by drawing on a wide variety of data sources we gained insight into present and future needs as well as those that are likely to persist across time. We also learned about needs at varying levels of granularity, and we identified factors that affect user behavior.

The needs analysis is based on data collected using multiple methods. We found that each source of data provided unique insights. Further, the combination of data from multiple sources helped to elucidate needs in particular areas. Users were, of course, an important source of information about their needs. TeraGrid personnel were also very knowledgeable, especially user consultants and support staff. These individuals were able to discern larger categories of user needs from the individual issues they dealt with on a daily basis. In addition, since they often had many years of experience, they were able to provide a perspective on user needs over time. Interviews with cyberinfrastructure experts and non-users of TeraGrid resources were informative, too. Finally, reports from domain-specific workshops organized by funding agencies and research communities helped to support and confirm much of what we learned through interviews.

Our results show that in order to gain a comprehensive understanding of user needs it is necessary to consider multiple factors. The term *ecosystem* has been used to describe the multiple and inter-related aspects of the world of high-performance computing.

> Supercomputing is not only about technologies, metrics, and economics; it is also about the people, organizations, and institutions that are key to the further progress of

> these technologies and about the complex web that connects people, organizations, products, and technologies (Graham, Snir, & Patterson, 2005, p. 157).

We identified four categories that are important to the user component of the broader TeraGrid ecosystem. First, we introduce the concept of *community of practice*, which we employ to help explain factors that affect the needs and behavior of current and target TeraGrid users. Second, we discuss overarching differences between current and target TeraGrid users. This simplified distinction is made in order to illustrate the marked division between users who belong to a community of practice that has a culture of high-performance computing with those who do not. Third, we analyze relationships between the nature of the research problems to be solved and TeraGrid's technical infrastructure, which is an important aspect of user needs. We end with an examination of potential communities of practice relevant to current TeraGrid users.

8.2.1 Communities of Practice and User Needs and Behavior

> Every one of those single users out there is very likely to be part of a larger community, depending on how you want to break it down. Do you just look at different scientific disciplines, or do you look at different types of scientific calculations, or..? Somehow, you can group these individual users. —TeraGrid participant

As the above quote illustrates, there are multiple ways to classify users. There are three main advantages to grouping users based on their needs, and the value and importance of doing so increases as the user population grows and diversifies. The benefits include

- improving strategies for delivering TeraGrid support and consultation services to users;
- informing decisions related to hardware, software, and policy; and
- providing mechanisms for users to interact with each other in support of their needs.

Our results show that the concept of *community of practice* (CoP) is a useful framework to understand and even anticipate the needs of users. Classifying users according to the communities of practice to which they belong provides an avenue for TeraGrid to interact with particular groups to design, develop, and implement strategies for technical infrastructure, user support, and education, outreach, and training that best meet their needs. Below we briefly discuss the community of practice concept. Following this, we discuss TeraGrid-relevant CoPs based on findings from the user needs analysis.

Community of Practice

A community of practice is defined as a group of people who share an interest in a domain of knowledge and who develop a set of approaches that allow them to deal with that domain successfully (Lave, 1988; Wenger, 1998). In this case, the term *domain* refers to a shared competence that distinguishes members from other people. A domain could be a group of biologists working on a similar problem or a network of educators exploring online learning. A CoP shares technology, language, culture, and common ways of addressing recurring problems. Participation in the community provides opportunities for learning that aids an individual to become a member of the group. The CoP framework helps to account for

variations in the needs of current TeraGrid users and provides a way to assess the needs of groups that are not yet using TeraGrid. Individuals belong to multiple communities of practice; they may be full participants in some and peripheral members of another.

We identified several communities of practice based on our analysis of user needs. We present these findings in two parts. First, we discuss general distinctions between current TeraGrid users and target users based on the Unified Theory of Acceptance and Use of Technology that we employed in the *TeraGrid User Survey*. Second, we analyze communities of practice relative to current TeraGrid users.

8.2.2 Users and the Unified Theory of Acceptance and Use of Technology

The Unified Theory of Acceptance and Use of Technology (UTAUT) is described in detail in Part 2 of this report. To summarize, according to UTAUT, four constructs play a significant role as direct determinants of user acceptance and usage of information technology. The two most important constructs are performance expectancy and effort expectancy. We describe each construct below and briefly discuss its relevance to current and target TeraGrid users. The analysis of current users is based primarily on results from the *TeraGrid User Survey*, and the main source information on target users is interviews with gateway developers (see Zimmerman & Finholt, 2007).

Users and Performance Expectancy

Performance expectancy is the degree to which an individual believes that using TeraGrid will help him or her to attain gains in job performance, and it is an important factor in whether or not a "culture of HPC" exists for particular groups of individuals. An alternate name for the concept of performance expectancy is *usefulness*. Results from the *TeraGrid User Survey* show that individuals who currently utilize TeraGrid resources do so because they are critical tools for their research. For example, 91% of respondents either strongly agreed or agreed that supercomputing is necessary to answer research questions of interest to them, and this was reinforced in interviews and workshops. This necessity is reflected in a statement made by a TeraGrid staff member.

> I've always felt that with supercomputing, you almost have to have desperate people. Because nobody in their right mind would use this if you have the luxury. You have to be really in a bad way to use shared resources.

At the other end of the spectrum are researchers who have not yet begun to conceive a need for TeraGrid. Prior research has shown that the major determinant of technology use and adoption is performance expectancy. Before new users can make judgments about the expected performance of a system, they must understand how it can help them to accomplish work.

Users and Effort Expectancy

Effort expectancy is the degree of ease associated with use of the system. While current users have a bias in favor of the usefulness of TeraGrid, they are favorable to a somewhat lesser degree in terms of its ease of use. Slightly fewer than half of the respondents indicated that

TeraGrid is easy to use, and this category contained a larger percentage of neutral responses than other items on the survey. Interviews with TeraGrid personnel and current users confirmed findings regarding ease of use. Thus, the lack of ease of using supercomputers, particularly shared resources such as TeraGrid is an issue for both current and potential users. The difference is that some people will tolerate usability barriers if the technology is important to them. For example, the director of a campus computing center said, in reference to users with large allocations:

> They will jump through hoops of fire to get their cycles. You can make the lives of those guys almost impossible, and they are going to get their resources.

On the other hand, if the system is not perceived as being useful, users will not be drawn to it no matter how easy it is to use.

Users and Facilitating Conditions and Social Influence

Facilitating conditions relate to an individual's belief that an organizational and technical infrastructure exists to support use of the system, and *social influence* is the degree to which an individual perceives that important others believe he or she should use the system. For many reasons, it is likely to be the case that individuals who do not perceive a use for TeraGrid will not be members of a community that has conditions or people that support and promote use of the technology. For example, in most cases, the problem to be solved on a supercomputer is derived from a mathematical model of the physical world. However, some areas of study do not yet lend themselves to such quantification. For example, in the humanities and in some social sciences models do not exist to study the vast digital archives of text, audio, and visual data that are becoming available. Given this, it is not surprising that historians and literary scholars are not typical users of HPC and that they do not have the trained personnel, algorithms, or applications to support their use. In addition, as they become users of HPC resources, their use of the resources and the systems required to support their use are likely to be much different from traditional HPC users.

The purpose of the above analysis has been to show that the needs of current and target users are, in general, quite different and should be addressed separately. We have analyzed the needs of target users elsewhere (Zimmerman, 2007; Zimmerman & Finholt, 2007). In the remainder of this section, we focus on issues that impact the needs of current TeraGrid users.

8.2.3 Relationship between Research and Technical Infrastructure

> We really need to understand what it is that they do—do they transfer a lot of data? Do they do a lot of computation? Can they benefit from use of multiple resources? Is visualization important to them? What are all of those things? —TeraGrid participant

One of the most common ways to assess the needs of users is to investigate the relationship between the nature of the research problem to be solved through the use of TeraGrid resources and technical requirements such as machine architecture, network infrastructure, system software, data storage and management, and tools for collaboration. The nature of the research problem affects the utilization of resources such as the "expense" of the calculation

and the ease of applying computation to the problem.[37] There are different ways to classify problems such as linear or nonlinear; applied versus theoretical; experiment versus simulation; by scale; and by the ratio of resources needed to invest in the storage of data versus the compute elements−to name a few examples. The alignment between the nature of the problem to be solved and various aspects of the technical infrastructure is an important and complex part of satisfying user needs. This helps to explain why it was a major topic raised in the interviews and user workshops we conducted and in the CI reports and other documents we examined. Below we introduce the technological components that are most often mentioned. Since the relationships between research problems and technical infrastructure are complex and have been covered extensively in other documents we provide only a brief introduction and simple examples.

- Machine Architecture

A supercomputer is composed of processors, memory, I/O system, and an interconnection network. Different configurations of these components are more or less effective for particular types of problems. For example, researchers who are trying to compare very large datasets prefer machines with large memory and high I/O. If the amount of data in the simulation is small, huge amounts of memory are not the concern they are for cases in which there are large files of data. Researchers who start with a fundamental theory and little data generally need compute power.

- System Software

System software includes the operating system components as well as tools such as compilers, schedulers, run-time libraries, monitoring software, debuggers, file systems, and visualization tools (Graham, Snir, & Patterson, 2005, p. 159). System software is not linked as directly to the nature of the research as other parts described here, although users may be grouped according to the compilers they use, for example. It is unclear how useful such categorizations are, however. Regardless, we mention system software because it is an important part of the HPC technical environment. The impact of system software is related more to usability and ease of use. For example, users must know how to find the parallel libraries they need and how to use the commands for compiling and running programs on each TeraGrid resource. One of the most important aspects of system software is the queues and the policies that control them.

- Network Infrastructure

We refer here to the high-speed network that connects TeraGrid sites. The network integrates high-performance computers, data resources and tools, and experimental facilities around the country. The high-bandwidth network is critical, for example, to users who compute across multiple sites, move large amounts of data from one place to another, or wish to improve computational models that connect with experiments.

[37] Graham, Snir, and Patterson provide an overview of supercomputing applications in a dozen domains, which is similar to our concept of the nature of the research problem.

- Data Storage and Management

Systems and tools to manage and store data are an important part of the technical infrastructure for many users. TeraGrid, for example, provides users with access to long-term archival storage systems and to TeraGrid-based collections; software to connect data resources over a network; and centralized file systems for short- and long-term data storage. User needs will vary depending on the amount of data to be managed, where the data reside (for example, in many distributed files or a few large files), the heterogeneity of the data, and analysis needs such as direct comparison of real-time data to simulations or cross correlation of massive data sets−to name a few examples.

- Collaboration Tools

The mention of collaboration tools in the context of HPC is relatively new. There are multiple needs for collaborative technologies such as helping researchers to connect with others who have expertise they require and sharing knowledge among widely dispersed users. For instance, some interviewees who conduct large-scale simulations told us that they would like a way to locate collaborators who are experimentalists in order to help them validate and improve their simulations.

The main purpose of the above was to introduce an important component of user needs. In some cases, communities of practice emerge from the relationship between the nature of the research and technical infrastructure requirements. For example, those who simulate processes at the same length scales (e.g., nanometers) or conduct atomistic simulations using codes such as AMBER or GROMACS may form CoPs. Below, we describe other potential CoPs that emerged from the needs analysis we conducted that may be useful for supporting and interacting with users.

8.2.4 Communities of Practice and Current TeraGrid Users

In some areas represented by users who currently utilize TeraGrid, broad disciplinary areas are helpful to understand users, but in many cases, it is difficult to generalize based on discipline. Results from the user survey, for example, did not find a significant difference between discipline and perceived usefulness of TeraGrid. We hypothesized that this is due to the fact that those who use TeraGrid are more like each other than they are different because HPC resources are vital to their research.[38] In addition, we observed during the TeraGrid planning workshops that even within groups, people from nominally similar fields often had very different computational needs for their research (see Lawrence & Zimmerman, 2007a, pp. 9-11). Additional evidence to support these findings is available in many of the cyberinfrastructure reports we described earlier. For example, the authors of a document on CI needs in the mathematical and physical sciences stated:

> Because of differing computational needs and complexity implications across disciplines, a common and important theme emphasized that one size does not fit all−across MPS, different [NSF] divisions and groups within divisions will have

[38] Note that we refer only to people who use TeraGrid. We cannot generalize about disciplinary areas because we did not compare those who use TeraGrid with those who do not. This would be informative, but it was beyond the scope of this study.

> different needs in order to access cutting edge science in their fields (MPSAC Working Group, 2004, p. 7.)

In addition to diversity within disciplinary areas, it is also common for individual users in the same project to employ multiple codes and to have different compute requirements based on the scale or type of problem they are investigating at a particular moment. As one user told us, "Our scientific interests are quite broad, and it's not easy to make classifications." Another said, "This is a big group, and we've got several different focuses."

Findings from the survey show that individuals who use TeraGrid more frequently are also greater users of TeraGrid support, more strongly identified as TeraGrid users, perceive themselves as more experienced, and are more positive about TeraGrid's usefulness, ease of use, and the facilitating conditions for using TeraGrid. Further, as one would expect, more frequent use is associated with larger allocation size. By itself, however, allocation size is not a sufficient way to categorize users because of the diversity of research problems addressed, software used, etc. An analysis of data we collected suggests more fruitful classifications, including experience level; algorithms, models, and software; and mode of usage. Communities of practice based on these dimensions may have greater potential in terms of devising strategies to identify and support user needs.

Experience Level

Individuals who are new to TeraGrid and to HPC face many similar challenges. TeraGrid is a complex system and *use* of TeraGrid requires multiple steps and specialized knowledge. First, there is a set of practical tasks that one must complete in order to use TeraGrid. For example, users must write a successful proposal to receive an allocation on a TeraGrid resource or work with someone who has an allocation. Once they have access to an allocation and to a user name and password, applied knowledge is needed in order to, for instance, select the most appropriate resource on which to run their application, figure out which versions of software and applications reside on particular machines, run jobs and retrieve data, and avoid data loss. A user's needs, experiences, and behavior are influenced by the level of knowledge and skill they have to negotiate these and other steps and the options they have for gaining the expertise they require. If users are part of a larger project in a CoP with a culture of HPC, they are likely to have access to help locally. However, this is not always the case, and even when it is PIs and their team members benefit from access to help outside their project (see Zimmerman & Finholt, 2006). A TeraGrid user support person explained how users' HPC background affects their experience with TeraGrid.

> In many ways a lot of the people that started doing science on the TeraGrid initially had already done all this stuff elsewhere. Either they had allocations elsewhere, or they were using super computers somewhere else. So, they knew the drill, and they knew their way around. … But the more new people you add to this...and we're adding people who are using these systems who are not schooled in computer science. That's not what they do, but they need to crunch some numbers. So, we're running into people with that sort of background that have to jump through lots of hoops. If you look at the TeraGrid user documentation, there's a lot of it and it's very complex. If you have multiple accounts at multiple sites, there are multiple ways of

> getting on and it's very confusing. I would say that's probably a big challenge. The complexity just to get on and do their science is amazing.

This interviewee felt, though, that TeraGrid had made good progress in reducing the complexity, thereby making it easier for new users to get started.

The type of support needed by beginning users varies. For instance, individuals who are the first in their department or project to use TeraGrid have many start-up questions. The initial one generally concerns how to get a TeraGrid allocation. Development allocations are intended to help people learn about the resources and to prepare them to apply for a larger allocation in the future. The process to obtain a development allocation is much simpler than for larger allocations in that the written proposal can be very brief—as short as a paragraph—and the review process is less formal. Even so, navigating the process can be challenging. We asked an individual who was the first person in his department to attempt to use TeraGrid, to document his experience. The notes he took appear in Appendix A. The issues he encountered are similar to those described to us by others and include:

- Lack of an overview of the application process and an example of a successful proposal
- Questions about terminology and eligibility requirements
- Usability issues with the interface to the proposal system that made the preparation and submission process more confusing than necessary

Some people who are new to TeraGrid do not have to complete the proposal process because it has been done by someone else on their research team. For these users, questions concern how to get an account, log-in, submit jobs, and keep track of their allocations. One potential user we interviewed suggested that TeraGrid provide online support in the form of a discussion board, chat, or bulletin board system (BBS).

> At least one BBS should be designated to the beginner−How to start to use the TeraGrid. At that point, we don't care about discipline. But once you become a user then you might have more specific questions related to your discipline.

As the quote above indicates, as users get past the initial hurdles, new ones arise. While discipline may be useful in some cases to categorize users, we found others to be more promising.

Algorithms, Models, and Software

TeraGrid users often employ the same algorithms, models, or software to study systems in similar as well as quite different subject areas. The types of codes used have implications for user support and for the future.

Results from the *TeraGrid User Survey* showed statistically significant differences in the codes that disciplines use. Specifically, biologists and chemists are more likely to use third-party codes or third-party codes augmented with some of their own routines/libraries. Computer scientists and astronomers favor codes developed entirely by themselves or their

group. Geoscientists use codes developed by their group and augmented with third-party software. On the other hand, Monte Carlo methods are a class of computational algorithms used to simulate many physical and mathematical systems, and models of fluid flow are necessary in astrophysics, aircraft design, climate modeling, and geophysics (Graham, Snir, & Patterson, p. 73). Philosopher Paul Humphreys (2002) has suggested organizing science based on computational models. According to Humphreys, *computational templates* are at the heart of computational models.

> Familiar examples of such templates are differential equation types, such as Laplace's equation and the Lotka-Volterra equations; statistical models such as the Poisson process and its various extensions; and specifically computational models such as cellular automata and spin-glass models (Humphreys, 2002, p. S2).

Besides the use of the same or similar algorithms and models, TeraGrid users in different subject areas frequently employ the same codes to model physical systems. For example, NAMD is a parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems that is used by many of those we interviewed. AMBER is another widely used code for molecular dynamics simulations. A geochemist described how similar codes are used to solve problems in multiple fields.

> It only makes sense to use petascale capability for really large problems that require massively parallel computations. This is a challenge by itself to make software run efficiently on this scale. Luckily for us, there's a much larger community of life science and biophysics people who are doing molecular simulations for protein folding and things like that. Essentially, we are using the software and approaches they develop to solve our problems of geochemical and environmental chemical processes.

The ability to use codes developed by others studying similar problems was of particular value to this MRAC user since he was one of the few people in his field using HPC resources.

The type of software used can affect the needs of users in terms of the support they require. For example, third-party software can simplify the lives of users, but it may increase the knowledge required by support personnel. For example, many chemistry researchers use pre-packaged software such as Gaussian. One of the TeraGrid interviewees mentioned that their site has specialists who deal specifically with chemistry users because they are not what he described as "sophisticated programmers. Further, he said:

> They just want to know: "How do I solve this problem?" And they want to stay on chemistry. And they're able to do that pretty well. They don't have to learn as much about computers. Physicists generally tend to be really nuts and bolts. They're really good programmers. They get right down to the lowest possible level. Those are kind of stereotypes, but they're pretty accurate, I think.

Another of the TeraGrid support personnel we interviewed observed changes in the kinds of questions he received between the early days of TeraGrid and the time we interviewed him in summer 2006.

> Early on, it was really oriented towards people who had a directory full of source code, and they were going to go out and build their own program. They were collaborating with researchers at another site or a national lab or whatever, and they would compile their own program, and they didn't need any third party software. As more and more people come in you start seeing different types of users.

He described how these changes impacted user support.

> It changes the type of question. You have to be familiar with the actual application software and maybe have people on staff. Going from telling people how to build programs and not really have to worry about what the program was doing—whether it was a physics program or ten molecular dynamics simulations for someone in chemical engineering. Compilers are compliers; command line options are command line options, and here are the libraries. Most of the users that we dealt with were running their own code as opposed to commercial code.

A support person at another RP site did not perceive significant differences in user needs over time, but this may have been due to the nature of the two sites−he was located at one of the original NSF-funded supercomputing sites and the other person was at a site newer to providing national level support.

The type of software they use also has implications for users. Thos who use third-party codes may not require as much expertise in terms of programming and development, but they also have less control over the direction of the code. For example, commercial software is proprietary and cannot generally be modified by users, even if they have the skill to do so. Open-source software is modifiable, but to do so requires expertise. In addition, variations made to a code by individuals or groups affects the degree to which it can be shared among different groups or used to replicate results. Individuals and groups that develop their own codes have more latitude to make changes to increase the scalability of the code, for instance, but this requires a significant investment of time and resources, particularly if the code is large. These issues will affect users into the future as we discuss later.

Although it is possible to make some generalizations regarding software used by particular domains, other data show that caution is required. For one, the needs of users in the same discipline can vary. An above quote describes physicists as skilled programmers. This may be true generally, particularly in comparison with fields such as chemistry, but physics researchers told us they have diverse needs that are influenced by the different scales with which they work (Lawrence & Zimmerman, 2007a, p. 9). Second, some users use more than one code. For example, a high-energy physicist we interviewed stated that “some kinds of measurements involve certain kinds of software and some platforms work better with other kinds of software.” Finally, while the user survey asked respondents about the *most*

important type of code related to their research goals, generalizable data on other aspects of software use (e.g., number, names, and origin of applications used) are lacking.

Usage Mode

Users who are employing new usage modes such as co-scheduling or real-time computing may benefit from interactions among themselves and with TeraGrid. Users could share lessons learned with each other and work with TeraGrid as a group on issues that will enhance their work while eliminating redundancy in interactions.

The areas described above, along with others that should be determined through further investigation and collaboration with users, provide avenues for supporting and interacting with a large and diverse user base. Communities of practice are likely to evolve over time with existing ones losing relevance and new ones being developed. For example, as a mode of usage becomes common, the benefit of classifying users in this way may diminish. Or, a CoP formed around those with an interest in petascale computing may be useful while researchers are working to adapt codes to this new environment, but it may become less relevant as these challenges are overcome.

8.3 User Behavior: Consequences of Unmet User Needs

As users try to meet their research needs relative to the use of TeraGrid resources, they exhibit behaviors that have implications for technology acquisition, RP policies and procedures, and the progress of science. We discuss two areas that affect user behavior that stood out in our analyses: queue times and usability issues.

8.3.1 The Problem of the Queue

Users are united in their dissatisfaction and frustration with job turnaround times. This was the most commonly mentioned barrier to TeraGrid use according to findings from the user survey and interviews with users and TeraGrid personnel. While it is a difficult problem to solve for technical and policy reasons, lengthy queue times have consequences that go beyond user inconvenience and irritation. Our findings show that long waits in the queue affect the speed at which science is conducted, deter users from making optimal use of resources, and increase the time spent on computing-related tasks versus "doing science."

Non-Optimal Use of Available Resources

Long queue times discourage optimal use of HPC resources by individuals who are not currently TeraGrid users as well as researchers who have allocations on the TeraGrid. One consequence of lengthy waits in the queue is that researchers expend resources to purchase, manage, store, and supply power to local clusters (e.g., project or departmental) that might be better utilized on institutional, state, regional, or national resources. It is not unusual for projects to maintain their own clusters, and this solution makes sense in some circumstances. One of the main advantages of researcher-owned clusters is control over the resources. Besides not having to wait in the queue, researchers can gain direct access to the cluster to install software, reconfigure systems, or troubleshoot problems. Local resources also provide students and researchers with opportunities to learn. Researchers who employ a combination of resources—for example, local, regional, and national—are able to submit jobs to different

resources based on factors such as fit to the task and priority. Thus, it is not a simple matter to judge the merits of one approach over another. It is all the case that choices about where and how to compute are not always based on an analysis of the best approach in a particular circumstance. As one scientist, who does not use TeraGrid resources, said:

> We've found that running our own resources and running them in the way we want to and configuring them in the way that we would like has been… Well, that's the way we've chosen to do it. Whether it's been more efficacious it's hard to say.

There is evidence, however, that in the future it may become more difficult for researchers to rely solely on project or departmental resources. For example, a faculty member in computational fluid dynamics, who primarily used the cluster he purchased with Department of Defense (DOD) funds, noted that three-dimensional (3-D) simulations take a long time and utilize a lot of processors. Although he had received funds from DOD to purchase a new cluster, he stated that the system he was purchasing would be inadequate for 3-D simulations. Thus, he grudgingly obtained a TeraGrid allocation to support this work, although it was used primarily by his doctoral students.

Earlier, we observed that it can be challenging to determine appropriate use of TeraGrid resources (see section 8.1.3). Long waits in the queue are one reason for this. Several researchers stated that they will request less processors than they need if it will reduce their job turnaround time. As one said:

> If you have to wait three times as long to get twice as many processors, you get more science done if you run on a smaller set. It's all the economics that go with a large, shared resource.

This quote leads to a related problem with overly long job turnaround times. As one user said, "I'm most interested in performance: How much can I get done with a certain amount of wall clock time?" Thus, getting more science done often translates into time spent devising strategies to "beat the queue."

Beating the Queue vs. "Doing Science"

Users, especially those who have the skills to do so, spend significant amounts of time designing and executing strategies to get through the queue faster. Sometimes this comes at the expense of the most efficient use of computational and human resources. As one user related:

> Right now on the Cray at Pittsburgh, my highest throughput is if I just run one hour jobs. So, I submit lots of short, one-hour jobs. I go through the queue much faster than I would if I did eight hour chunks, but that's eight times the work of managing all the different stuff. It's not really eight times because I have it automated, but… We're always finding tricks as to how to get into the machines as quickly as possible.

We heard many other stories in which users described how they tried to speed up the turnaround time of their jobs. Besides spending time trying to get through the queue users

also lose time doing science when their job does not complete for some reason. When this happens they have to get back in line and wait again for their job to run. According to interviews with TeraGrid personnel and users, wait times of over one week are not unusual. Thus, the time lost due to a job failure can be significant. This is one reason that users, particularly those who are trying to increase the scalability of their code, want debug queues that match the scale they are trying to achieve.

> The debug queue is good, but a lot of time what we're trying to do is increase the scalability of our code. “Need this many processors, this much memory for a short period of time.” Trying to do scalability parallel program development really limits how big your science can get, how fast you can get work done.

The size of the job can have different affects on users’ productivity. As one interviewee said, “If you've got a short job that two-day wait in the queue is a lot more significant than if you've got a 12-hour job or a large job.”

Interestingly, long queue times can contribute to the development of more portable codes since they may encourage researchers to spend time tuning and optimizing their code to run on multiple architectures, particularly ones that fewer people are able to utilize. For instance, the developers of AMBER, a widely used molecular dynamics code, have worked to make the code run on a variety of architectures. This benefits users who employ this code. An individual in a group that had developed their own code described a similar strategy.

> We applied for allocations from different machines to compare performance and to distribute work more evenly because some computers are more crowded. It's an optimization problem between the CPU time spent on solving the problem and human time spent on waiting in line in the queue to get the job run.

Again, it is generally more sophisticated users who can take advantage of these strategies. Even these users often prefer to compute on a limited number of machines because “if you have a strange result, you want to know that everything ran on this machine and nothing changed.”

Clearly, many aspects of user behavior are related to attempts to get as much research done as possible in the shortest amount of time. Finding ways to speed up the turnaround time of jobs is one way that users try to accomplish this. Another factor that limits the productivity of researchers is problems with the usability of tools and interfaces.

8.3.2 Usability

Usability refers to ease of use, effectiveness, efficiency, and user satisfaction with tools and interfaces. Usable systems also have a low error rate, and when users do encounter errors, they are able to recover from them quickly. We identified several types of usability barriers that are common across user types based on open-ended responses to the user survey and interviews with users and TeraGrid personnel:

- Challenges to finding documentation and information on the TeraGrid web site
- Limitations of available tools for carrying out tasks such as submitting and monitoring jobs, tracking allocation balances, moving files, and transferring data
- Problems related to the management of login and account information
- Difficulty finding resources with desired applications and system software

Heterogeneity of resources and policies across RP sites exacerbated many usability challenges. Appendix D in Part 2 of this report contains representative responses from the open-ended survey item regarding barriers, many of which are usability problems, to TeraGrid use.

Usability issues affect users of all types and experience levels, although the particular issues vary based on experience level and other factors that we discussed previously. Still, we were surprised to find that advanced users with large allocations are interested in gateway-type access to resources, tools, and information. Interviewees gave two primary reasons for this. One is their desire for tools to manage their workflow, including submitting jobs and transferring data.

> What I would really like to see ultimately are workflow tools that allow us to manage where our data was, where our runs were because I've got 30 projects or something that are ongoing. If I had ways with electronic notebook where I could say, "Oh yeah, I ran this on the Cray the other day. I want to transfer it over to TeraGrid NCSA or TeraGrid San Diego. And this one's running there." I would also like ways to manage where the information is. Part of the solution is something like electronic notebooks and part of it is portals to the machines.

This LRAC user noted that he had just received his password for the TeraGrid User Portal, and he was hopeful that this would help him with workflow management.[39] A physicist with substantial HPC expertise whose project uses millions of service units each year spoke to the problem of managing the vast amounts of data that will be generated by new detectors soon to come online.

> If you can keep all your data on one disk to work on this stuff, who cares? Nobody cares. But if you have two petabytes of data per year, believe me, it will be a lot of guys who are sitting behind the scenes somewhere. … And a portal is a good way to make analysis as transparent as possible for the end user.

The interviewee stated that it already requires a minimum of two people to manage the utilization of their annual allocation on TeraGrid and other resources: one person who is deeply involved with the physics and a second person who specializes in technical aspects

[39] The TeraGrid User Portal (TGUP) is a Web interface for managing account status, for obtaining information about TeraGrid resources, and for accessing many of the existing TeraGrid services in a single place. See http://www.teragrid.org/userinfo/portal.php. The TGUP was in development during most of our study, so assessing user satisfaction with the portal was not part of our investigation. However, it does not yet appear to have the workflow capabilities this user was interested in.

such as data production, job submission systems, development, and tracking allocations on different clusters. If the latter set of tasks could be simplified it would free up time and personnel for scientific work.

The second reason that large users expressed interest in portals is to facilitate collaboration with experimentalists. Some researchers who conduct simulations are interested in collaborating with experimentalists as a way to improve their models. Conversely, experimentalists stated that they can benefit from computational guidance to help them determine what kinds of experiments to conduct. Modelers realize that experimentalists who are not HPC users will find HPC a poor fit with their practices; gateways are a means to make computing more accessible to some experimentalists and engineers who otherwise "wouldn't touch any of this computing with a ten foot pole."

While they share some things in common, overall, the needs of "power users" and new users who access resources via science gateways are likely to be different. For example, advanced HPC users are used to interacting directly with the resources and may find that a gateway interface, in many respects, does not offer the flexibility they want. On the other hand, the hope is that the gateway will enable these researchers to do things that were not possible or easy to do before. Referring to these types of users, one of the gateway developers described his project's vision in this regard.

> The researchers are capable of using them just as they are, and in fact that's what they are used to. What we're trying to do is give them another way—actually using grid mechanisms to get to the resources and give them different tools to use their codes and pull in data and do their work that they couldn't do before.

These efforts may move the grid vision forward, but achieving a stable and reliable system given all the different layers (e.g., research codes, grid software, gateway interface, TeraGrid resources) is a difficult technical challenge.

Some usability issues are easier to improve than others. Problems with existing interfaces such as the online proposal system (see Appendix A) can be fixed relatively quickly by employing accepted interface design and usability assessment techniques described in key texts (e.g. Nielsen, 1993; Nielsen & Mack, 1994; Shneiderman & Plaisant, 2005). The development of new tools such as those to manage workflows requires significantly more time and effort. Some of the science gateway projects are focusing on workflows, and this may be an area ripe for collaboration among gateways and between TeraGrid, gateways, and software developers.

8.4 Impact of TeraGrid: Results of Meeting User Needs

In earlier parts of this report we analyzed the impact of TeraGrid from the point of view of TeraGrid personnel and, to a lesser extent, from the perspective of the CI experts we interviewed (see section 6.4). We discussed successes to date as well as the longer term vision for TeraGrid. We also made some comparisons between user needs and TeraGrid services and development efforts. In this section, we present an analysis of users' responses to

interview questions concerning the impact of TeraGrid and/or HPC on their research. One question asked user workshop participants and interviewees to describe how TeraGrid and/or HPC resources and capabilities had impacted their work, knowledge, or scientific understanding. Second, we inquired about the benchmarks relevant to their field of science in terms of assessing impact (e.g., publication of results, faster time to solution, or increased scale of simulation).

TeraGrid and the individual RP sites play a significant role in the research of current users by virtue of the fact that the resources and services they provide are necessary to the work of these researchers as shown by findings from the user survey, interviews, and user workshop. Given the necessity of access, we were interested to go a step further and hear from users about specific ways to measure the impact of TeraGrid on their research. We noted previously that few projects are using TeraGrid to run at more than one site concurrently. For those that do, TeraGrid enables research that could not be done otherwise. For example, in order to simulate the entire arterial tree, George Karniadakis' team requires more shared memory than any one machine has available. By adapting his code to reduce latency, he has been able to run single jobs at multiple sites and thereby obtain the memory he needs to work toward his research goal.[40] Most users, though, continue to employ TeraGrid in traditional ways, although some users are taking advantage of the distributed nature of TeraGrid to move data from one place to another. Thus, we are just beginning to realize the unique ways in which TeraGrid will impact research outcomes; we describe some of the current benefits below.

Not surprisingly, users tended to emphasize impacts that are directly linked to research activities or outcomes such as

- publications, especially quality as measured by the prestige or impact of the outlet;
- reducing time to solution;
- making it possible to simulate phenomenon at longer time scales or across a continuum as resources become more powerful;
- influencing the direction of work in other areas or focusing the problem space; and
- facilitating collaborations.

One result of the last item is that it enables researchers to study processes that had been missed until the resources became available to simulate them. A high-energy theoretical physicist stated that as compute capability has increased their work has begun to have an impact on the analysis of experimental data, which has facilitated their collaboration with experimentalists and furthered discovery in both realms. Engineers noted that HPC has led to better designs.

> If you can improve the design process and enhance your confidence, then perhaps you can get by with fewer tests, or you design more rapidly, or you can fix something more rapidly. If something does go wrong, you'll know fairly quickly how to fix it.

[40] This is how the work was described to us by TeraGrid personnel, but it is consistent with the way the research of Karniadakis and his group has been publicized elsewhere.

This PI also stated, as did other interviewees, that jobs he assigns for homework now would have been research problems in the past. Another researcher described recent work by his doctoral student that took a few days, "whereas 5-10 years ago, it would have taken his entire PhD time-frame of anywhere from 4-5 years just to do that work."

Since HPC is a critical tool to current users, anything that reduces the time spent on computational tasks or makes them easier to accomplish contributes to the conduct of research. In the section on usability, we described some of the barriers that get in the way of "doing science." Although there are currently more usability challenges than solutions, several interviewees expressed sentiments similar to this LRAC chemistry user, who described the steps TeraGrid has made in tying together distributed resources and the positive affect it has had on his group's research.

> The main benefit for us of having them all connected in this kind of TeraGrid idea has been...well, probably several things, but the things I think of right now are the ability to quickly move huge data sets between different computers and things on this backbone. That has been invaluable when you need to move things around and so forth. If you want to use a number of different resources for the same project then to have to transfer those via normal things would be very tedious. … They have one thing they call the TeraGrid cluster that's really meant to be kind of the same environment everywhere even though it covers five or so different sites. And it's the same architecture, and it's the same everything, so you can almost just move your binary in between these machines, and there's no porting cost, and you submit it via the same way and everything. We're starting to see some of that benefit… And so as far as the DOE goes, you don't have any of that. So, you've got the full cost every time you go to use a different site at DOE; you have start from scratch to get it to work there.

Capturing descriptions of the research impact of TeraGrid such as the one above are necessary and important means of assessment as they reveal outcomes that would otherwise be difficult or impossible to capture. In addition, the impact of a particular technology is hard to predict because users often employ tools in ways that designers did not expect; interviews with users, TeraGrid personnel, and others are one means to understand changes that are developing or in process. Quantitative methods such as citation counts or the impact factor of journals that publish work based on the use of TeraGrid resources are useful, but they are inadequate by themselves (TeraGrid Impact RAT, 2006).

Some interviewees also spoke to future impacts in the sense of the longer term vision for TeraGrid. Their views on ultimate success aligned with the sentiments expressed by TeraGrid and CI experts. They believed that TeraGrid success should be assessed in terms of its role in enabling changes in the practice of science and engineering; its contribution to the development of usable tools, technologies, and software to enhance research productivity; and its ability to expand the communities of users through gateways. In the next section, we analyze the future needs of users to gain additional insight into the ways TeraGrid might impact research.

8.5 User Needs: Looking Ahead

The user needs analysis was structured to collect data to help identify and understand needs for the future as well as the present. In the previous section, we focused on the ways that TeraGrid and HPC have impacted the work of researchers based on two questions we posed to interviewees and workshop attendees. A third and related question asked individuals to focus on the future and to describe the computational, social, and/or organizational factors that constrain their research productivity or hinder their ability to address research questions they would like to answer. Our findings relative to the future needs of users coincide with topics discussed in cyberinfrastructure reports and elsewhere, including the final report from the TeraGrid Planning Process (Killeen et al., 2008). Since current needs often foreshadow future issues, we have already discussed most of these areas directly or indirectly elsewhere in this report. These include needs related to:

- managing, storing, and analyzing growing amounts of data
- parallelization of codes to, for example, simulate processes on longer time scales, or to make existing codes run efficiently on HPC resources
- modes of usage (e.g. on-demand, real-time) that accommodate a wider range of needs
- education, training, and support
- tools and support for collaboration

The number and capability of TeraGrid resources is increasing along with the number of users. For example, we noted earlier that a petascale system is scheduled to come online in 2011. Although it is not currently slated to be integrated into TeraGrid it will be of interest to TeraGrid users who require lots of processors with high-performance interconnects. We also mentioned NSF's Track 2 initiative, which 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). Of course, as new resources come online existing systems reach the end of their usable life and are retired. Still, the overall capacity and capability of the resources will increase. Will these new resources lead to reduced wait times and better accommodate the needs of different types of users? Unfortunately, a positive answer to this question seems unlikely under the current scenario. For one, most of the large users we interviewed told that us that they adjust the size of the research problem to the size of the machine that is available. As one user said:

> In our program we reach the limit of the machine before we reach the scale of the problem we are truly interested in. So we always adjust the size of the problem that we are working on to the size of the machine that we have available."

A primary need for these users is access to more cycles for longer periods of time. Second, the move to petascale will affect the needs of users and present many technical challenges. For example, as a TeraGrid participant noted: The “output from the petascale has to go into a terascale machine to do some filtering, massaging, and maybe only then you'll be able to take it home and make sense of it.” Dealing with the amount of data generated by a petascale computer is only one challenge. Significant effort and resources will be needed to make software run efficiently at this scale. In addition, the petascale system will be suitable to

particular type of problems. These and other issues concerning petascale computing have been discussed elsewhere (e.g., Snavely; Yeung et al., 2007).

8.6 Summary: TeraGrid User Needs

Meeting the needs of TeraGrid users presents several challenges. For one, the application of user-centered design methods to a large, diverse, distributed, and growing population of users is a complex and resource-intensive process (Spencer et al., 2006; Zimmerman & Nardi, 2006). Specifically, the design, and delivery of a cyberinfrastructure such as TeraGrid is difficult because the users are

- numerous,
- widely distributed, and
- include a heterogeneous mix of users whose needs and priorities may conflict and who differ in terms of culture, skills, knowledge, and other factors.

We conducted an analysis of user needs—the first step in the UCD cycle. The purpose of the analysis was to gain an understanding of the factors that influence the needs of users, affect their decisions about how or whether to use TeraGrid, and help to explain variations in their needs. There are many types of research problems that can benefit from TeraGrid resources, and they require varying combinations of architecture, software, policy, and support. We found the concept of community of practice (CoP) to be useful framework to help classify users, and we identified several CoPs, which have the potential to assist TeraGrid to devise strategies to further study, interact with, and support the needs of its users.

Ideally, each step of the UCD process—needs analysis, design, development, and system implementation—is iterative and ongoing. The resources and expertise that this would require adds to the challenges of meeting the needs of TeraGrid users. However, employing the full UCD cycle would help to increase the usability of TeraGrid tools and systems, which our results show is important to all types of users. Improving the usability of TeraGrid tools and systems would enable users to spend more time on the conduct of science and less on computing-related tasks. The conduct of science is also hindered by job turnaround time, which limits the productivity of current users and deters potential users from employing TeraGrid resources.

In order to effectively utilize information on user needs collected in this study and that which appears in other sources such as cyberinfrastructure reports (see section 8) TeraGrid must find ways to address dilemmas that result from:

- the demand to support more users with limited resources;
- limitations in the methods used to assess the impact of TeraGrid; and
- the lack of clarity regarding what constitutes appropriate use of TeraGrid resources.

Science gateways are one means to attract and support TeraGrid users. Users may also be important sources of support for each other. For example, popular codes and applications generally have their own web sites, tutorials, and mail lists. TeraGrid does not need to

duplicate these online resources but could link to them as a service to users. In addition, TeraGrid personnel might monitor these lists or have their own online mechanisms (chat, wiki, listserv, etc.) for TeraGrid-specific questions related to the use of particular computational models, applications, or codes.

In addition to the challenges of supporting the present needs of current users, TeraGrid must also be attuned to needs that will arise in the near future. Meeting future—and present—needs will require collaboration across TeraGrid sites as well as collaboration between TeraGrid and users, software developers, science gateways, educators, and others. Although it has active programs in most of these areas, the issues below seem to be outside the scope of TeraGrid to address on its own.

- Developing human capacity at the undergraduate and graduate levels in computational science
- Broadening participation whether it is by attracting users from minority-serving institutions, new disciplines, etc.
- Developing or adapting existing codes to run efficiently on TeraGrid resources or scaling of codes and algorithms to take advantage of new resources
- Integrating CI across a wide set of resources providers (e.g., national, international, campus) to provide pathways for users

Developing and implementing a coordinated approach to these and other issues that concern all stakeholders is an open challenge. The responsibility for addressing it is beyond the expertise or resources of any one organization.

Finally, users react to unintended as well as purposeful incentives in their use of computing resources. More thought should be given to the behaviors to be encouraged and to strategies that will motivate users to exhibit preferred behavior. This will not be easy given the many different needs to be met. It is also unlikely that issues and challenges related to the use of a shared system will ever be entirely alleviated. Several interviewees noted that telescopes and colliders are managed by the communities they serve. Some computing resources are similar to this, too. In TeraGrid, however, no single user community controls the resources. Each approach should be examined more carefully to gain insight into its benefits and disadvantages.

9. TeraGrid Science Gateways

Science gateways were described at the beginning of this report (see section 3.1.3) and have been mentioned in a number of places in the text. In this section, we analyze science gateways in more depth. The goal of gateways is to enable entire communities of users associated with a shared research goal to use TeraGrid resources through a common interface. Science gateway projects are similar in that they have external funding to build a community-specific cyberinfrastructure; many of them pre-date TeraGrid. Although a few of the projects receive funding from TeraGrid, this is not their primary source of support. There were approximately twenty projects designated as TeraGrid Science Gateways when we

began the evaluation study.[41] By the time of this report, the number had grown to almost thirty-five.[42] A wide range of disciplines are represented including astronomy, biology, chemistry, computer science, earth science, engineering, materials science, nanotechnology, and physics.

It is difficult to describe gateways along dimensions such as purpose, governance, permanence, source of funding, capabilities, and audiences because there is substantial variation among them in these respects. On its web site, TeraGrid describes three common forms of gateways.

- A gateway that is packaged as a web portal with users in front and TeraGrid services in back.
- A gateway that involves application programs running on users' machines (i.e. workstations and desktops) and accesses services in TeraGrid (and elsewhere).
- A gateway that bridges multiple grids, allowing communities to utilize both community developed grids and TeraGrid.

Most of the gateways we studied are funded by NSF as time-limited collaborative research and development projects similar to those described by Lawrence (2006), but others are supported by a combination of funding sources or are embedded in programs or institutions that have longer term stability and more formal and ongoing interactions with their intended user communities.

Examples of the kinds of capabilities that gateways are developing include the ability to run complex climate simulations, to query large databases, or to simplify the submission of jobs to supercomputer resources. For example, Linked Environments for Atmospheric Discovery (LEAD) is attempting to bring together meteorological data, forecast models, and analysis and visualization tools to explore the weather as it evolves. LEAD's goal is to automate many of the time consuming and complicated tasks associated with meteorological science. The developers of LEAD are trying to serve a range of users from scientists who are experienced in modeling and simulation using HPC resources to school children and everyone in between. Not all the projects we studied are designing for such varied users. A report from a workshop conducted in June 2007 as part of the TeraGrid planning process provides an overview of the purpose, funding sources, target user communities, and status of most (n=17) of the science gateway projects that existed at that time (Lawrence & Zimmerman, 2007b).

In spite of differences in the characteristics of gateway projects and the approaches they take to achieve their objective, we identified two broad goals that are similar across gateway projects. First, gateways aim to support new types of science and to enable the pursuit of novel research questions.

[41] A project is designated as a TeraGrid Science Gateway if it has an allocation on the TeraGrid. Working with gateway projects, TeraGrid developed a community allocation whose goal is to delegate account management, accounting, certificate management, and user support to the gateway developers.
[42] The TeraGrid web site includes a section on the Science Gateways program. See: http://www.teragrid.org/programs/sci_gateways/

> So, the whole goal, at least in my opinion, is to enable these scientists who are not accustomed to these big machines to start using them. Once that's done, what one would expect is that they will start asking qualitatively bigger, more complicated questions. So, it will be a self-fulfilling process where eventually, they will start asking questions so big that meta-computing starts to look like a path to solution.

This is similar to what other interviewees told us. A second shared goal among the gateway projects we studied is to make the technology invisible. Hiding the technology makes it possible for users to concentrate on doing science as the quote below illustrates.

> It should be a black box. All they should worry about is the decisions that are relative to the scope of their science. … Where are my data sets? What kind of calculations? 'What if' type questions.

Or, as another interviewee stated, "It should be become very transparent. It should become like a power grid." As we noted previously, increasing the usability of TeraGrid would be beneficial to all types of users. Thus, gateways have the potential to be valuable to experienced as well as new users.

Developers of science gateways have two roles: 1) TeraGrid user, and 2) intermediaries between user communities. Our analysis of these roles draws on interviews conducted as part of the evaluation study and the June 2007 planning process workshop described above. Since we have published these results in two other documents (Lawrence & Zimmerman, 2007b; Zimmerman & Finholt, 2007), we summarize the findings in the sections below and refer interested readers to the more complete reports.

9.1 Gateway Developers as TeraGrid Users

As TeraGrid users, developers of science gateways are in need of things that help make development easier and that assist them to support their users. We found that gateway developers are excited by the potential of TeraGrid to make HPC available to end users and communities who would otherwise be unable to conduct their research as effectively or efficiently. In addition, they are enthusiastic about the opportunity for distributed communities to work together on common solutions. Meanwhile, they are eager to move TeraGrid toward a collaborative mindset that enables the developers to focus on the unique needs of their gateway communities. At present, they find that too much energy is focused on re-creating custom solutions when standardized systems or a TeraGrid-hosted gateway layer would suffice. Specifically, science gateway developers have need for:

- basic services that gateways can use instead of creating or hosting their own;
- templates and standardized systems to save developers the time of recreating things that others have already built; and
- standardization that would make TeraGrid a *real* grid that could support the effective use of allocations and meta-scheduling.

They would also like to find ways to operate more effectively as a community in order to better support education and development needs of gateway developers. This is similar to the need for collaboration support mentioned by some individual users of TeraGrid.

9.2 Gateway Developers as Intermediaries

It is well known from prior research that in order for users to adopt new technologies, they must offer advantages over current practices, positively change the way that work can be performed, and be easy to implement and straightforward to use (Star & Ruhleder, 1996; Venkatesh et al., 2003). This is also recognized by the gateway projects as succinctly stated by a PI of one the gateway projects we studied: "Unless there is something extra that a scientist can get, they won't adopt any new technology."

Science gateways are a type of mediating organization, and they may play a key role in attracting new users to TeraGrid.[43] Specifically, they provide important social and technical support that help new users to conceptualize a use for TeraGrid and increase their willingness and ability to use it.

An interesting aspect of gateways is that most of the requirements do not come from the intended users. This is because many potential users do not yet perceive a need for HPC resources and capabilities. As one interviewee succinctly stated, "It's hard to sell something to people where the expectation is zero." Thus, c*onceptualization of use* is concerned with helping potential users to see a relationship between the research questions they want to answer and the capabilities of TeraGrid. *Willingness to use* TeraGrid relates to how TeraGrid fits with users' values, expectations, and practices. For example, TeraGrid has made adjustments to policies and procedures such as user authentication and the tracking of allocations to accommodate the needs of gateways. Both conceptualization of use and willingness to use are largely anticipatory activities and depend on gateways' and/or TeraGrid's abilities to foresee barriers and to work to reduce or eliminate them. Finally, in order for individuals to use TeraGrid resources through gateways, they must know *how to use it* to submit jobs to remote resources, for example, or transfer data across sites. The usability of gateway interfaces to TeraGrid resources and tools is an important aspect of use.

Gateways also play an important role in facilitating interaction between TeraGrid and new communities of users. It is difficult for TeraGrid, which has less than 150 FTEs, to interact with potentially thousands of new users. Thus, the gateway concept can also be viewed as a mechanism of interaction. The results from our study of TeraGrid Science Gateways show that attracting new users to TeraGrid often involves intense and ongoing activities on the part of both the gateways and TeraGrid, but that gateways bear much of the difficult task of helping potential users to conceptualize a use for TeraGrid.

[43] Most gateways are too new to have yet made significant use of TeraGrid. Thus, it is not known if widespread use of TeraGrid via science gateways will occur.

9.3 Summary: Science Gateways

Science gateway developers comprise an important group of TeraGrid users because of the various ways in which they build on and enhance the capabilities of TeraGrid. As users, they require services that make it easier for them to develop tools and services that meet the needs of their user communities.

The developers of science gateways mediate between the needs of potential users and TeraGrid. It appears that they may play an important role in introducing TeraGrid to potential users and supporting the use of these communities in ways that fit with their culture, expectations, and skills. Gateways may also benefit current users, and some projects are focused primarily on this goal. For one, gateways may help reduce the barriers to grid computing. Our interviews with current users of TeraGrid show that their needs would also be served by the work that the gateways and TeraGrid are doing to make the TeraGrid infrastructure more transparent, and thus, easier to use. TeraGrid users comprise a broad spectrum, however, and a small percentage of users currently use the majority of TeraGrid's resources. While these users might welcome improvements in ease of use, it is not their overriding concern. The prime issue for "hero users" is to obtain as much of the available resource as they can, and they will put up with a lot of pain on the usability side to achieve this goal. More than once in the interviews we conducted with TeraGrid personnel and individual users we heard that the "high-end, high-performance user is going to do whatever they've been doing." As one person put it, "It's hard to teach old dogs new tricks."

Most gateways are funded as limited-time research projects. This situation has important implications for the stability and sustainability of these organizations and for TeraGrid's ability to rely on the roles they fulfill (for example, see Ribes & Finholt, 2007). In addition, if TeraGrid provides support to gateways or if funding agencies decide to extend funding for these projects, then there must be ways to identify and evaluate factors related to success of science gateways—first in attracting new users and later in finding ways to sustain and adapt to that use. Finally, at this point in time, many gateways are not necessary to users. Interviewees noted that databases, such as the Protein Data Bank are resources that scientists in certain fields require. In many respects, TeraGrid is also necessary because it has unique capabilities and resources. Gateways are generally not in this advantageous situation. They must work hard to cultivate users for both themselves and for TeraGrid.

10. Discussion

The primary purposes of the TeraGrid evaluation research project were to conduct an analysis of the

- needs of TeraGrid users, including TeraGrid Science Gateways,
- the impact of TeraGrid on user’s work, and
- the relationship among the TeraGrid partners.

We also conducted two surveys to assess the satisfaction of those who attended tutorials held at the TeraGrid conferences in 2006 and 2007. We hope the results from the study will provide useful feedback to TeraGrid and NSF that will help with future planning and

program improvement. It is also our intent that the findings contribute to the literature on virtual organizations and the evolution of HPC in the United States and to approaches to analyze the needs of users who are distributed, heterogeneous, and numerous—properties that are characteristic of e-Science.

TeraGrid is part of a potentially major shift that is underway in the delivery of high-performance computing resources and services supported by the NSF and in the institutions, policies, technologies, and users that are part of this *socio-technical ecosystem*, the phrase used to describe the complex web that connects people, organizations, products, and technologies (Graham, Snir, & Patterson, 2005, p. 157). Resource providers are independent, but also linked together, forming a virtual, networked organization characterized by distributed and dynamic governance and coordination processes. By studying TeraGrid, we have learned about the specific tensions that arise when computing resources and support are delivered through a VO. We were able to study how the tensions we analyzed were approached, managed, or avoided and to understand the factors that lead to particular actions on non-actions. For example, TeraGrid is now packaging common software in kits. Resource providers are required to install a basic package, but installation of other components is left up to each RP. This solution balances site autonomy while providing users with some standardization across sites. Achieving a clear vision in the light of continual change has been more difficult for TeraGrid to reconcile. However, since most personnel have a similar view of what it will mean for TeraGrid to succeed in the long run, the organization may be able to build on this to clarify shorter term goals.

The vision for HPC is that it will evolve into *cyberinfrastructure* that brings together distributed resources such as computational tools and services, instruments, data, and people to accelerate the pace of science and engineering discoveries (e.g., Hey & Trefethen, 2005; NSF, 2007). The research areas and problems that require HPC are expanding, portending a dramatic rise in the number and types of HPC users. The needs of these new users have already begun to influence resource provider policies and practices, technology development, and education, training, and outreach programs. Meeting the needs of these users may also help the larger population of individuals that use TeraGrid. Like the power grid, the vision of *HPC as infrastructure* means that it will be transparent to users. Currently, however, using a supercomputer is not a simple matter of "signing on and hooking up" (Star & Ruhleder, 1996). Making high-end computing resources such as TeraGrid easier to use is seen as necessary to increase research productivity and speed up knowledge production (e.g., Nomura, 2005; West, 2007). Ease of use does not have to come at the expense of capabilities and power, however, as noted by Donald Norman, a popular product design consultant (2008, n.p.).

> Everyone wants simplicity. Everyone misses the point. Simplicity is not the goal. We do not wish to give up the power and flexibility of our technologies. … People want the extra power that increased features bring to a product, but they intensely dislike the complexity that results. Is this a paradox? Not necessarily. Complexity can be managed. … The real issue is about design: designing things that have the power required for the job while maintaining understandability, the feeling of control, and the pleasure of accomplishment.

Good design is relevant beyond the technical interface; it applies to processes such as obtaining allocations, getting a TeraGrid account, and making arrangements to conduct runs at multiple sites. The science gateways share the goal to enable new types of science. For many gateways, this means developing ways to make it easier to obtain an allocation, sign on, and automatically select resources on which to run. Usability is an outcome that would benefit all users because it would allow them to spend more time on science.

10.1 Limitations of the Study

The main limitations of the investigation arise from the complexity, dynamism, and scope of the TeraGrid virtual organization, the number and diversity of current users, the lack of previous studies to guide some data collection and analysis activities, and the evolving state of HPC, including the challenge of meeting the needs of new communities of users. We discuss the impact of these challenges on the investigation and the ways in which we attempted to mitigate them.

We conducted formal interviews with individuals located at five of the nine RP sites that were part of TeraGrid during our study. Given the diversity of the individual sites, there may be important perspectives that are not represented in this report. We attempted to reduce the impact of this limitation by selecting sites with different characteristics and by attending and observing meetings where representatives from all nine sites were present. We recognize that public statements made by project participants may differ from views that would be expressed in private interviews. We are encouraged by the fact that common themes emerged from the interviews conducted with individuals across the sites and with different roles in the organization.

TeraGrid is a large and complex project that grew out of and exists in complex environment that includes past collaborations and competitions and changing policies, technologies, and needs for HPC in science and engineering. This limited the investigation in three primary ways. First, it was beyond the scope of the study to conduct a comprehensive examination of the historical, political, social, and technological landscape related to TeraGrid. Our findings are based largely on what interviewees told us and what we observed. We were not able to compare this data with close examinations of funding solicitations, the larger NSF portfolio of HPC resources, or the detailed history prior to TeraGrid, particularly the PACI program. Second, while we were provided with excellent access to people, meetings, and documents, there were many internal TeraGrid conversations, discussions between NSF and TeraGrid, and interactions between TeraGrid and external parties to which we were not privy. Third, there are important stakeholders such as middleware developers and other providers of HPC and grid resources that play a role in the TeraGrid ecosystem; in-depth study of these stakeholders was beyond the scope of this investigation and its objectives.

Assessing the needs of current and target TeraGrid users posed several challenges. First, TeraGrid counts more than 4,000 individuals among its present users—a number that does

not include most people who access TeraGrid resources through science gateways.[44] As mentioned elsewhere in this report, these users are diverse along several dimensions. Since this was the first comprehensive scientific study of an HPC user population that we are aware of, there was little prior research on which to base our investigation. By combining qualitative (interviews, user workshop, participant observation) and quantitative (survey) methods, we gained a generalizable picture of current users as well as in-depth information that helps to explain the statistical findings. Further, we employed a theoretical framework that has been tested in numerous studies of technology adoption to guide the design of our survey. Second, although we studied all types of individual TeraGrid users, we focused on users affiliated with a project that had a large resource allocation because these individuals utilize the majority of TeraGrid resources. Again, we employed multiple research methods to collect data from a variety of sources to help reduce concerns that results might be skewed toward the needs of LRAC users. Finally, we investigated the needs of target TeraGrid users primarily through interviews with science gateway developers, who play a key role in attracting new users to TeraGrid and supporting their use. As new communities of users employ TeraGrid, it will be important to assess their needs directly.

10.2 Future Research

The history and evolution of NSF-supported HPC systems and networks, including study of the policies that have shaped the environment over time; the nature of institutions that provide the resources and support their use; the characteristics of users and the factors that affect their needs; and the impact of HPC on scientific outcomes have received little consideration from scholars in history or social science. This is not to say that these topics have been completely overlooked (e.g. Aspray & Williams, 1984; Rogers, 1998), but the attention has been minor relative to the importance of HPC on science and engineering research and the investment in high-end computing in the United States and elsewhere. This knowledge is necessary to evaluate options for resource delivery, develop curriculum in computer science and in the domains, and assess user needs and develop systems and policies to meet those needs. The consequences of these gaps in our understanding will only increase going forward. The research areas and problems that require HPC to address questions of interest are expanding, portending a dramatic rise in the number and types of HPC users. At the same, petascale computing offers unprecedented capability for researchers able to capitalize on this power, but it also presents many challenges. Adapting codes to the petascale environment, managing and analyzing the data produced, and developing the human capacity to both support and use a petascale computer are some of the problems to be faced. As this report shows, the technical challenges are difficult, but the social, institutional, and organizational challenges of effectively and efficiently enabling researchers to do their science using distributed resources and services are equally difficult. These issues are not unique to TeraGrid, but are relevant to e-science in general. Below, we discuss some of the most pressing needs for research raised by the TeraGrid evaluation study.

The results presented in this report suggest two lines of research. First, attention should be directed to ongoing and long-term investigation of research areas addressed in the evaluation,

[44] This is the latest figure we have.

especially those related to the TeraGrid virtual organization and to the needs of current and target TeraGrid users. It can take a long time for outcomes to occur and to become visible. For example, we are just beginning to understand how users are employing capabilities that TeraGrid has developed and how TeraGrid has dealt with the many tensions inherent in a VO. Second, the study's findings also point to topics to investigate in greater depth.

Although data collection for the evaluation was limited to one year, we gained insight into the evolution of the TeraGrid partnership over time. This was possible because many of the TeraGrid personnel we interviewed have been involved in the project since the beginning of their institution's participation. Many users and cyberinfrastructure experts also had a long-term perspective on HPC and TeraGrid. TeraGrid continues to be dynamic, and entirely new processes and issues are likely to surface based on changes in composition of the TeraGrid partners and the portfolio of resources. For example, as our investigation was ending, TeraGrid was implementing changes to its governance structure to accommodate growth. What issues were these revisions intended to address, and are they achieving the desired goals? Effective mechanisms of coordination and communication are critical to virtual organizations. The evaluation identified the main approaches used by TeraGrid, but detailed study of the use of collaborative technologies and shared sources of data and information would help us to better understand their role in providing coordination and cohesion across the distributed organization. Finally, a survey of TeraGrid personnel would provide generalizable information that would contribute to the literature on virtual organizations and could be used to enhance the program.

Findings from this study provide guidance for future surveys of the TeraGrid user population. For instance, a survey designed to better understand relevant communities of practice would inform mechanisms to interact with and support the needs of the user community. This information would also be useful in designing virtual communities to help users support their own needs. In addition to generalizable information gained through surveys, there is a need for in-depth study of a small sample of users drawn based on the characteristics found to influence user needs and behavior such as discipline, scale of investigation, allocation level, codes and algorithms used, and level of experience. Case studies would allow us to test and refine what we learned in this study, surface other factors that affect user needs and behavior, help in devising policies and strategies that motivate and incentivize user behavior to better serve the entire community, and inform development to meet future needs such as managing and analyzing vast amounts of data.

Finally, science gateways are one approach being used to anticipate and deal with challenges related to increasing both the total number of users and the types of disciplinary communities that employ TeraGrid. More research is needed on science gateways as mediating organizations and on the needs of the target user communities. Since most gateway users are not routinely accessing TeraGrid resources through the gateway it will be important to conduct additional user analyses as new communities utilize TeraGrid. It will also be useful to identify and evaluate factors related to success of science gateways, first in attracting new users, and later in finding ways to sustain and adapt to that use.

11. Acknowledgments

We thank TeraGrid personnel, especially Charlie Catlett and Dane Skow, former directors of the TeraGrid Grid Infrastructure Group, for facilitating this research. We also acknowledge the many individuals who contributed to this research through surveys, interviews, and workshops; the work reported here would not have been possible without their participation. This report is based upon work supported by the National Science Foundation under Grant No. OCI 0603525 to the University of Michigan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

12. References

Aspray, W., & Williams, B. O. (1984). Arming American scientists: NSF and the provision of scientific computing facilities for universities, 1950-1973. *Annals of the History of Computing, IEEE 16*(4), 60-74.

Berman, F., Fox, G., & Hey, T., eds. (2003). *Grid Computing: Making the Global Infrastructure a Reality*. Wiley, New York.

Borgman, C. L., Wallis, J. C., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. *International Journal on Digital Libraries 7*(1-2), 17-30.

Bryson, J. M. (1995). *Strategic Planning for Public and Nonprofit Organizations: A Guide to Strengthening and Sustaining Organizational Achievement*. Jossey-Bass, San Francisco, CA.

Catlett, C., Beckman, P., Skow, D., & Foster, I. (2006). Creating and operating national-scale cyberinrastructure services. *CTWatch Quarterly 2*(2), 35-41.

Catlett, C., Goasguen, S. & Cobb, J. (2006). *TeraGrid Policy Management Framework*. TG-1: Policy.

Catlett, C., et al. (2008). TeraGrid: Analysis of organization, system architecture, and middleware enabling new types of applications. In L.Grandinetti (Ed.), *High Performance Computing (HPC) and Grids in Action.* Advances in Parallel Computing, Volume 16. IOS Press, Amsterdam.

Collins, H. M. (2003). LIGO becomes big science. *Historical Studies in the Physical and Biological Sciences 33*(2), 261-297.

Creswell, J. W. (2003). *Research Design:Qualitative, Quantitative, and Mixed Methods Approaches*. Sage, Thousand Oaks, CA.

Cummings, J., Finholt, T., Foster, I., Kesselman, C., & Lawrence, K. A. (2008). *Beyond Being There: A Blueprint for Advancing the Design, Development, and Evaluation of*

Virtual Organizations. Retrieved Augusts 21, 2008, from http://www.ci.uchicago.edu/events/VirtOrg2008/VO_report.pdf

Cummings, J. N. & Kiesler, S. (2005). Collaborative research across disciplinary and institutional boundaries. *Social Studies of Science* 35(5), 703-722.

Cummings, J. N. & Kiesler, S. (2007). Coordination costs and project outcomes in multi-university collaborations. *Research Policy* 36, 1620-1634.

DeSanctis, G., & Monge, P. (1999). Comunication processes for virtual organizations. *Organization Science*, 10(6), 693-703.

Dubrow, A. (2008). Testing the limits of the standard model. Retrieved July 31, 2008, from http://www.tacc.utexas.edu/research/users/features/dynamic.php?m_b_c=sugar

Edwards, P., Jackson, S., Bowker, G., & Knobel, C. (2007). *Understanding Infrastructure: Dynamics, Tensions, and Design. Report of a Workshop on History & Theory of Infrastructure: Lessons for New Scientific Cyberinfrastructures*. University of Michigan, School of Information, Ann Arbor, MI.

Foster, I., & Kesselman, C. (2004). *The Grid: Blueprint for a New Computing Infrastructure* (2nd ed.). Elsevier, Boston.

Graham, S. L., Snir, M., & Patterson, C. A. (Eds.). (2005). *Getting up to Speed: The Future of Supercomputing*. National Academies Press, Washington, DC.

Hacker, T., & Wheeler, B. (2007). Making research cyberinfrastructure a strategic choice. *EDUCAUSE Quarterly 30*(1), 21-29.

Hey, T., & Trefethen, A. E. (2008). E-Science, cyberinfrastructure, and scholarly communication. In G. M. Olson, A. Zimmerman, & N. Bos (Eds.) *Scientific collaboration on the Internet*, pp. 15-31. MIT Press, Cambridge, MA.

Hey T., & Trefethen, A. E. (2005). Cyberinfrastructure for e-science. *Science*, *308*(5723): 817-821.

Humphreys, P. (2002). Computational models. *Philosophy of Science 69*(Supplement 3), S1-S11.

Killeen, T. L., et al. (2008). *The Next Generation Research Grid: A Path Forward. Final Report.* Retrieved August 24, 2008, from http://teragridfuture.org

Krause, M., & 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 Academies of Sciences Washington, DC.

Lave, J. (1988). *Cognition in Practice: Minds, Mathematics and Culture in Everyday Life*. Cambridge University Press, Cambridge, UK.

Lawrence, K. A. (2006). Walking the tightrope: The balancing acts of a large e-Research project. *Computer Supported Cooperative Work 15*, 385-411.

Lawrence, K. A., & 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 July 11, 2008, from http://www.teragridfuture.org/system/files/TeraGrid+User+Workshops+Final+Report.pdf

Lawrence, K. A., & 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 July 11, 2008, from http://www.teragridfuture.org/system/files/TeraGrid+Science+Gateways+Workshop+Report.pdf

Lee, C. P., Dourish, P., & Mark, G. (2006). The human infrastructure of cyberinfrastructure. Pages 483-492 in *Proceedings of the Conference on Computer Supported Cooperative Work* (CSCW '06). ACM Press, New York.

Lindgaard, G., Dillon, R., Trbovich, P., White, R., Fernandes, G., Lundahl, S., & Pinnamaneni, A. (2006). User needs analysis and requirements engineering: Theory and practice. *Interacting with Computers* 18, 47-70.

Monteiro, E. (1998). Scaling information infrastructure: The case of next-generation IP in the Internet. *The Information Society 14*(3), 229-245.

MPSAC Working Group. (2007). *Identifying Major Scientific Challenges in the Mathematical and Physical Sciences and Their Cyberinfrastructure Needs.* A workshop funded by the National Science Foundation held on April 21, 2004. Retrieved July 7, 2008, from http://www.nsf.gov/attachments/100811/public/CyberscienceFinal4.pdf

National Science Foundation. (2001). *Distributed Terascale Facility (DTF), Program Solicitation (NSF 01-51)*. National Science Foundation. Retrieved March 30, 2008 from http://www.nsf.gov/pubs/2001/nsf0151/nsf0151.htm

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 for 21st Century Discovery*. National Science Foundation, Arlington, VA. Retrieved July 7, 2008, from http://www.nsf.gov/pubs/2007/nsf0728/index.jsp

Nielsen, J. (1993). *Usability Engineering*. Academic Press, Boston.

Nielsen, J., & Mack, R. L. (1994). *Usability Inspection Methods*. John Wiley & Sons, New York.

Nomura, M. (2005). Trends in high-end computing in the United States government. *Science & Technology Trends Quarterly Review 16*, 46-60.

Norman, D. (2008). Simplicity is not the answer. Retrieved July 31, 2008 from, http://www.jnd.org/dn.mss/simplicity_is_not_th.html.

Norman, D. A., & Draper, S. W., eds. (1986). *User Centered System Design: New Perspectives on Human-Computer Interaction*. Lawrence Erlbaum, Hillsdale, NJ.

Olson, G., Finholt, T., & Teasley, S. (2000). Behavioral aspects of collaboratories. In S. Koslow & M. Huerta (Eds.), *Electronic Collaboration in Science*, pp. 1-14. Lawrence Erlbaum, Mahwah, NJ.

Olson, J. S., et al. (2008). A theory of remote scientific collaboration. In G. M. Olson, A. Zimmerman, & N. Bos (Eds.) *Scientific collaboration on the Internet*, pp. 73-98. MIT Press, Cambridge, MA.

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 Millennium*.

Ribes, D., & Bowker, G. C. (2008). Organizing for multidisciplinary collaboration: The case of the Geosciences Network. In G. M. Olson, A. Zimmerman, & N. Bos (Eds.) *Scientific collaboration on the Internet*, pp. 311-330. MIT Press, Cambridge, MA.

Ribes, D., & Finholt, T. A. (2007). Tensions across the scales: Planning infrastructure for the long-term. Pages 229-238 in *Proceedings of the International ACM SIGGROUP Conference on Supporting Group Work (GROUP '07)*, ACM Press, New York, NY.

Shneiderman, B., & Plaisant, C. (2005). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (4th ed.). Pearson, Boston.

Snavely, A., Jacobs, G., & Bader, D. A., eds. (2006). *Workshop Report: Petascale Computing in the Biological Sciences*. Retrieved July 9, 2008, from www.sdsc.edu/pmac/workshops/bio2006/pubs/PetascaleBIOworkshopreport.pdf

Spencer, B. F., Jr., Butler, R., Ricker, K., Marcusiu, D., Finholt, T., Foster, I., & Kessleman, C. (2008). NEESgrid: Lessons learned for future cyberinfrastructure development. In G. M. Olson, A. Zimmerman, & N. Bos (Eds.), *Science on the Internet*. MIT Press, Cambridge, MA.

Star, S. (1999). The ethnography of infrastructure. *American Behavioral Scientist 43*(3), 377-391.

Star, S. L., & Ruhleder, K. (1996). The ecology of infrastructure: problems in the implementation of large-scale information systems. *Information System Research 7*, 111-134

TeraGrid Impact Requirements Analysis Team. (2006). Measuring TeraGrid impact: Methods to document effects of TeraGrid resources and capabilities on scientific practice and outcomes. Retrieved July 7, 2008, from http://www.teragridforum.org/mediawiki/images/4/4a/Impact_RAT_report_final0906.doc

U.S. Department of Energy, Office of Science. (2007). *Allocation Procedure for Leadership Computing Facilities*. Retrieved August 19, 2008, from http://www.science.doe.gov/ascr/INCITE/ProposedAllocationProcess.pdf

Venkatesh, V., et al. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly* 27(3), 425-478.

Wenger, E. (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press, Cambridge, UK.

West, J. (2007). Supercomputing in the participation age. *HPC Wire* 16, 14. Retrieved July 31, 2008, from http://www.hpcwire.com/offthewire/17898754.html

Wilkins-Diehr, N. (2006). Special issue: Science Gateways—common community interfaces to grid resources. *Concurrency and Computation: Practice and Experience*, *19*(6), 743-749.

Yeung, P. K., Moser, R. D., Plesniak, M. W., Meneveau, C., Elghobashi, S., & Aidun, C. K. (2007). *Cyber-Fluid Dynamics: Final Report to the National Science Foundation on NSF Workshop on Cyber-Fluid Dynamics: New Frontiers in Research and Education*. Retrieved July 3, 2008 from http://www.nsf-cyberfluids.gatech.edu/cyberfd_finalreport.pdf

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

Zimmerman, A., & 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 March 30, 2008, from http://www.crew.umich.edu/research/teragrid_user_workshop.pdf

Zimmerman, A. and Nardi, B. A. (2006): 'Whither or whether HCI: Requirements analysis for multi-sited, multi-user cyberinfrastructures', Page 1601-1607 in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*, ACM Press, New York, NY.

Appendix A: Notes on Applying for a DAC Allocation

The notes below were created by an individual as he went through the process to obtain a DAC allocation. The individual is a PhD student and was the first person in his department to attempt to use TeraGrid. In order to protect his identity we do not provide the exact dates of his application except to say that he applied for the allocation in late 2007-early 2008.

Day 1:

1. Went to the main TeraGrid website
2. Data and computation resources. http://www.teragrid.org/userinfo/hardware/index.php
3. POPS page. Thought to myself they should have a numbered list of steps I need to go through to create a new account. https://pops-submit.teragrid.org/
4. Created a POPS login. My first password was too short. Didn't realize it needed to be a certain length. Will this login be my TeraGrid login? If I login using the different authentication mechanisms will it take me to different places? What is the difference between the portal and the website that I am using now?
4. Took me a little while to figure out that I needed to read the user guide. This guide is the list that I was looking for:
http://www.ci-partnership.org/Allocations/pops_guide.html#docs

--Logging in and creating the proposal

5. It is an annoyance to go through the steps in one window and create the proposal in another window. A pdf document with numbered steps that I can print out would have been nice.
6. After I have already created my POPS password I now see that my password needs to be a specific length.
7. In selecting the proposal type the language is not consistent with the resource allocation terminology I've seen before: startup, medium, large VS DRAC, MRAC, LRAC. I picked a startup allocation.
8. Upcoming meetings page. There are five options available. Not exactly sure which to choose. When do these different committees meet? Which is most relevant to my application? The different committees seem to be categorized by organization rather than scientific application.
I picked the TeraGrid DAC because I want to apply for a TeraGrid allocation.
9. While filling out the PI application, I began to wonder about the required qualifications to be a PI. I could be wasting my time. I entered my position as research assistant (which I technically am)

Day 2:

10. Logged in, opened 2 windows, one for the account procedure page and the other with the actual account creation page.
11. Selected *edit current proposal*.

12. Thought: Might have been nice to see a sample proposal.
13. Not sure if my abstract should be a paragraph or 5 pages.
14. Not exactly sure how many SUs I need. It's also in bold and red. Not sure why.
15. What is the difference between multi-site and cross site?

Day 3:

Had to update my curriculum vitae

Day 4:

The interface has changed. The changes were subtle enough that it took me a few minutes to realize that everything has changed. I just want to log in and continue with my application. I somehow came to this page: http://teragrid.org/userinfo/access/dac.php that says I can log in on the left column, but there no longer is a left column.

Finally found the POPS login page, but I seem to have forgotten my username. Remembered the password and username.
Having problems uploading my cv. I can find the document, but the document will not upload.

The status is showing incomplete.

For the heck of it I clicked the submit button, and the submission status changed to *submitted*. Not sure if my cv was successfully uploaded or not.

Update- just got an email saying that my application was submitted.