

TeraGrid Science Advisory Board
10-11 December 2009
NSF, Arlington, VA

Executive Summary

1. Broadening Participation

The TeraGrid has as part of its vision: a larger and more diverse community of STEM practitioners, within the constraint that the TeraGrid is not funded as a national resource. As such, broadening participation in high-performance computing is a goal, which is shared by the NSF as well as other federal agencies. One outcome of the Science Advisory Board's (SAB) discussion of broadening participation in computational science was that the TeraGrid and the NSF should consider a strategic planning effort, possibly in consultation with a sub-committee of the SAB, to map out a plan to realize the vision. The strategic planning process should (1) build on TeraGrid's strengths and grass-roots activities such as Campus Champions and Science Gateways, if such dependence on intermediaries is deemed a good strategy, (2) identify and overcome weaknesses such as gaps in nationwide coverage, (3) delineate activities that should and should not be undertaken by the TeraGrid, e.g. by clearly identifying potential leveraged assets provided by intermediaries, (4) clearly state the threats (e.g. stress on user support function and allocation process by growing community of users), (5) define the metrics of success, (6) engage the Advisory Committee for Cyberinfrastructure (ACCI), especially its task force on cyberlearning and workforce development, the NSF's Education and Human Resources directorate, and the NSF's program in Broadening Participation in Computing within the CISE directorate, and (7) quantify if and where resource limitations represent a barrier to success in this regard and assess potential solutions to overcoming such barriers.

2. Future of NSF-Sponsored HPC

Following a very wide-ranging discussion, the SAB made the statement that appears at the end of this summary.

3. Long-Term Data Storage

The SAB recommends that the TeraGrid go forward with plans to prototype long-term data storage systems, maintaining close coordination with the ACCI task force on data. The SAB also recommends that TeraGrid and XD encourage research groups to manage their storage more efficiently through various means including usage policies and information tools. The SAB supports, in concept, a data storage policy that requires raising the minimum metadata standards for any data in long-term storage.

The TeraGrid should also consider alternative organizational models and arrangements, including partnerships with existing data archives and service centers. The SAB noted that there is a long-standing empirical relationship between the computational capability that is available to researchers and the volume of data added to the long-term archive, which argues in favor of a balanced approach when attempting to assign priorities among HPC, data storage, software and broadening participation.

4. Sustaining a Persistent National Cyberinfrastructure

In a recent community workshop held by the ACCI HPC task force, the consensus that emerged was that cyberinfrastructure needs to be persistently supported, producing persistent growth in capabilities and maintaining user confidence in stability, which requires a long-term strategy with longer awards, a stronger focus on long-term data stewardship, a relatively small number of centers at the high end to take advantage of economies, capabilities and expertise concentration of scale. The SAB agrees in general with this assessment, but also recommends (1) evaluation of which aspects of HPC cyberinfrastructure need to be centralized and which can be distributed, and (2) determination of how best to address evolving user requirements within the lifetime of longer center awards.

5. Identifying, Building and Sustaining Strategic Partnerships

The SAB applauds initial efforts to build a partnership with the Open Science Grid (OSG). In addition, the SAB recommends that the TeraGrid take pains to communicate to users the benefits of a partnership with OSG, for example, enabling new activities, sharing computational load (moving high-throughput work from TeraGrid to OSG), and learning how to allocate resources more effectively (OSG has no allocation process at present). The SAB also suggests looking farther afield for organizations with which collaborations might be valuable.

6. Scientific Impact of the TeraGrid

The TeraGrid is often asked to measure, document and improve its scientific impact, but this is an acknowledged difficult problem. The SAB suggested:

- Recruiting highly-qualified reviewers for the allocation process (the SAB recognizes the difficulty in getting highly-qualified scientists to review infrastructure usage requests)
- Compiling “gems” or “highlights,” that is, exemplary papers from various disciplines. For example, the TeraGrid could include on its web site links provided by its users to papers (to the extent that such links do not violate copyrights) that acknowledge or have benefitted from TeraGrid support.
 - This list and/or the list from the TeraGrid annual report could be ranked, possibly by the SAB or TRAC panelists.
 - A group of experts (e.g., SAB members) could write periodic opinions on papers worth reading that were enabled by TeraGrid resources
- Continue its “Science Highlights” publication of science and engineering brief case reports useful for those not in the particular field, the general populace, students, administrators, and federal and state legislators
- Adopting additional metrics, e.g., citation index, impact index of papers acknowledging TeraGrid support; success of grant proposals citing prior TeraGrid usage; and
- Having TeraGrid publish its own journal.

7. Process for Selection of New SAB Members

The SAB recommends that the process of selecting new SAB members:

- Include input from the NSF Cyberinfrastructure Coordinating Committee (CICC) ;
- Include representatives of other agencies, especially those that fund TeraGrid users, or, at least, invite such representatives to attend SAB meetings;

- Consider increasing the size of the SAB to increase participation at each meeting;
- Establish a set of criteria for the population of the SAB that can be drawn from the following:
 - Knowledge about computing, familiarity with TeraGrid, but not just top 10 users of TeraGrid
 - Balance in gender and geographic distribution, especially EPSCoR states
 - Balance in disciplines, including those that aren't using "big iron" today
 - Representation from industry providing HPC, using HPC or offering complementary services (e.g. cloud computing)
 - Representation from underrepresented minorities, persons with disabilities or the institutions that serve them
 - Representatives from academic computing centers
 - Senior/experienced people who have looked at work being done in community, can serve as "representatives" of their discipline communities
 - Junior "rising stars" who are just starting out as HPC users
 - Experience with other computing environments, not just with TeraGrid
 - Representatives from other large programs such as PetaApps, SDCI, CDI, SciDAC or INCITE, or experience with similar programs outside of the U.S.

TeraGrid Science Advisory Board
Statement on the Future of NSF-Funded High-Performance Computing

“The NSF, via the open science TeraGrid and other, more discipline-oriented high-performance computing (HPC) facilities, supports fundamental computational science, based on merit review. The openness of the TeraGrid has been very successful, supporting computational research funded not only by the NSF, but also by several other federal and state agencies and, to a small degree, private companies. The TeraGrid Science Advisory Board (SAB) strongly supports this model of open access and strongly endorses continuing support by NSF for computational science research, both funded by NSF and funded by other agencies or entities. However, the SAB notes that the success of the open computational science model going forward depends on adequate resources being deployed for HPC facilities, including HPC systems and the software, networking and human expertise infrastructure that support them, which may be beyond the means of the NSF through its Office of Cyberinfrastructure (OCI). A gap already exists between the available HPC resources in the TeraGrid portfolio and the resources requested through the allocation process. This gap will be exacerbated by the retirement of several smaller systems in the near future, and the recent failure of the Track 2C award, contract and procurement.

Accordingly, the SAB encourages the TeraGrid to

- further develop a mitigation plan that (a) informs users of the existing and planned resources and the anticipated gaps, (b) provides a clear timeline of available resources, in terms of allocatable computational power, and their expected lifetime, and (c) helps users identify alternative resources (campuses, agencies) that may be suitable for their research.

The SAB encourages OCI to

- act quickly to remedy the immediate problem associated with the Track 2C failure and develop safeguards in the award and procurement process that help to prevent such failures in the future;
- work with the Cyberinfrastructure Coordinating Committee (CICC) of the NSF, and program directors in the relevant disciplinary divisions, to ensure that adequate computing resources are available and delivered to meritorious NSF-funded projects;
- work with the NSF Advisory Committee for Cyberinfrastructure in its task-force based strategic planning exercise to develop a long-term plan for delivery of high-quality, well-supported HPC resources; and
- reach out to other federal agencies, through the Office of Science and Technology Policy, or other interagency mechanisms, to develop a coordinated approach to support the Nation’s growing requirements for fundamental computational science.”

1. Introduction

The semi-annual meeting of the TeraGrid Science Advisory Board (SAB; see charge in Appendix A.1 and membership in Appendix A.2) was held at the National Science Foundation (NSF) in Arlington, VA on 10-11 December 2009. Over the course of the meeting (see agenda in Appendix A.3), several presentations of the current status of TeraGrid and pending issues were made. A lively discussion on several topics is summarized below.

2. Broadening Participation

The SAB took up the question of how to broaden participation specifically in TeraGrid, and more generally in computational science and high-performance computing (HPC). The NSF is committed to “broadening participation”, which has different, overlapping meanings in different contexts. In general, broadening participation refers to increasing the participation in STEM or specific science or engineering fields or endeavors, of women, underrepresented minorities and persons with disabilities, and the colleges and universities that specifically serve them. These are obvious dormant talent pools that can help meet the Nation’s growing demand for an educated STEM workforce, particularly in computing and computational science and engineering. This overlaps with the general development or education of the next generation of scientists and engineers. In general NSF parlance, the former is referred to as “broadening participation” and the latter is referred to as “education” or possibly “workforce development”. There is also the EPSCoR program cutting across all of NSF to diversify the participation of states so states that receive disproportionately less NSF funding can benefit from national programs. Lately, because of the emergence of, and potential for, new fields of computational science, engineering and other areas of scholarship, and the need to demonstrate the national breadth, scope and scale of a broad and open general science and engineering resource such as the TeraGrid, there has also been a desire to increase the number of users and the number of fields utilizing TeraGrid. Within current TeraGrid discussions and as presented to the SAB, all of these meanings fall within “broadening participation.” These aspects present multiple, different challenges for initiating and continuing the use of HPC. The TeraGrid, through its Education, Outreach and Training, as well as other activities, has striven to broaden participation. There has been a relatively recent confluence of efforts to better integrate the TeraGrid infrastructure and campus-based computing infrastructures, including human and technical infrastructures, through the Campus Champions program, and broaden participation, as this is one way to help reach a wider range of users, campuses and communities. While the TeraGrid is a national facility, increasing the pool of users with limited resources is challenging.

The discussion was initiated by Scott Lathrop of TeraGrid, who presented¹ a summary of TeraGrid education and outreach activities, the vision of which is more diverse generations of science, technology, engineering and mathematics (STEM) practitioners. Lathrop’s presentation highlighted efforts to involve underrepresented communities and education programs from middle school through 12th grade (7-12) and through undergraduate and post-graduate levels (13-20). The TG engages in a broad range of activities that include supercomputing education programs for high-school teachers and college faculty, summer workshops, “CI [cyberinfrastructure] days” on college campuses, small allocations to help get new users started,

¹ All presentations are available at http://teragridforum.org/mediawiki/index.php?title=Possible_Agenda_Items.

advanced support for knowledgeable users, and support for disciplinary communities through Science Gateways. Lathrop also highlighted the Campus Champions program, which currently has 54 member institutions with Campus Champions who volunteer their time to promote supercomputing on the TeraGrid at their home campuses, because it is widely viewed as a success. The Campus Champions program is currently conducted in collaboration with the Open Science Grid and the NSF Blue Waters project. Lathrop added that, while the Campus Champions program is actively working with a number of western states to address gaps in coverage across the country, TeraGrid reported in the last annual program plan that there are people in all 50 states who are involved in one or more TeraGrid programs, for example, through allocations of computing time and participation in education and training events. As with integrating any infrastructure, there are challenges in maintaining a common level of quality.

In the discussion that followed, a number of important points were raised:

- Based on input from a significant number of large TeraGrid users who also use other open science HPC resources, the TeraGrid Resource Providers (RP) should coordinate common CI issues (security, connectivity, data and computing interoperability, etc.) with supercomputing resource providers supported by other agencies, notably the Department of Energy (DOE) and the National Aeronautics and Space Administration (NASA). It was noted that extensive interoperability between RPs would be difficult due to institutional barriers and the significant costs involved. However, technical coordination on some basic capabilities, such as data transfers between these open science systems, could provide significant benefit at minimum expense.
- Some demographic information (e.g., belonging to a minority-serving institution or MSI) about those who request TeraGrid resources and those who review requests, should be collected and analyzed in conjunction with other TeraGrid user information such as results of satisfaction surveys, as has been done on occasion in the past.
- The TeraGrid collects data on participants that includes surveys of satisfaction and changes in practices that are requested after the first year of participation and every six months thereafter. Although it is also working to build a database to conduct longitudinal studies, a more strategic view of how to collect, use and leverage this type of information is needed, particularly with respect to information regarding new users and new fields.

Reaching under-represented communities is a matter of concern to many agencies and programs. TeraGrid should continue its partnership with such groups as the Minority-Serving Institution Cyberinfrastructure Empowerment Coalition (MSI-CIEC) and look into programs that have been established elsewhere, e.g., the Institute for Robotics and Intelligent Systems at the University of Southern California (<http://iris.usc.edu/>), as well as many others. Worthwhile partnerships could be developed with engineering and pre-engineering recruiting programs.

- In addition to the emphasis on training of new and current users, the TeraGrid should consider ways to create and enhance capacity, raising the sea level to support an ever-widening community. This requires a broad set of well-documented education activities that include
 - Determining competencies at K-20 levels, e.g., following up on research by the Ralph Regula School of Computational Science (<http://www.rscs.org/k12.shtml>) and others
 - Working with CASC, EDUCAUSE, Internet2 and InCommon, among others
 - Promoting computational science (simulation and modeling) competencies

- through the states' educational science education standards boards
 - Establishing summer institutes, possibly co-funded by the Computer Research Association's Committee on the Status of Women in Computing Research (<http://www.cra-w.org/>) or the Coalition to Diversify Computing (<http://www.cdc-computing.org/>)
 - Working with the National Education Technology Standards (NETS: <http://www.iste.org/AM/Template.cfm?Section=NETS>) effort
 - Proposing projects that are co-funded by the RPs' home states to develop standards for computational thinking and its relation to logical/quantitative reasoning
 - Linking to principal investigators of computational Major Research Infrastructure (MRI) projects and Science and Technology Centers (STC) on various campuses
 - Preparing how-to manuals on education, outreach and training best practices for state, academic and industry partners
 - Strengthening work with undergraduate, including community college, and graduate computational science and engineering curriculum and programs, in addition to the extra-curricular conference attendance, training and mentoring of students
- The group of users and potential users with disabilities is another population of concern to NSF and other agencies and should be served, at least not excluded by poor design or implementation. How are the needs of users with disabilities being met? How does the TeraGrid seek to improve the lives and productivity of people with disabilities beyond the research that is currently being done with sign language and Braille? While TeraGrid asks registrants of conferences and workshops if they have special needs, few requests have been received – awareness was identified as a particular concern. Several suggestions were made including
 - Offering closed captioning or taped copies for webinars;
 - Connecting to the NSF Learning Science Center's Virtual Language and Visual Language project (<http://vl2.gallaudet.edu/index.php>);
 - Identifying new ways of representing data, e.g., based on the work of Richard Ladner (<http://www.nvrc.org/content.aspx?page=31818§ion=8>); and
 - Developing a white paper entitled, "How TeraGrid Accommodates People with Disabilities."
- There is a serious concern with lack of resources. The tension between the desire to broaden participation, including the need to increase awareness of the TeraGrid, and the requirement to provide support for the influx of new users, especially those who are less experienced with HPC, is exacerbated by the limited resources available within the TeraGrid and the limited smaller-scale resources on campuses.

One upshot of the discussion was that TeraGrid should consider developing a strategic plan to realize the vision of *a larger and more diverse community of STEM practitioners* that builds on strengths (e.g., the Campus Champions and the Science Gateways), identifies and overcomes weaknesses (e.g., the voids on the U.S. map where supercomputing has not penetrated as thoroughly into the STEM research and education practices, notably concentrated in the EPSCoR states), and clearly articulates the threats (e.g., stress on user support function and allocation process by growing community of novice or near-novice users – the pyramid of support

metaphor in which the bulk of support is provided to least experienced users) and the metrics of success. The strategic plan should combine the grass-roots approach of Campus Champions with a more systematic organization, if such dependence on intermediaries is deemed a good strategy. The strategic planning process should deeply engage the Advisory Committee for Cyberinfrastructure (ACCI), especially its task force on cyberlearning and workforce development, the NSF's Education and Human Resources directorate, and programs for broadening participation located elsewhere in the NSF, such as the Broadening Participation in Computing (BPC) program within the Computer and Information Science and Engineering (CISE) directorate (http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=13510&org=CISE). For example, one of the results of the BPC program in CISE is that teaching the teachers is an effective strategy (also demonstrated in using modeling as a paradigm for physics instruction – see David Hestenes program at Arizona State Univ. - <http://modeling.asu.edu/>); this suggests that a similar strategy could be employed in computational science as is being promoted for computer science.

Much of the discussion focused on the fact that the TeraGrid is not funded to be a national resource. For that reason, any strategic planning must include determination of what activities the TeraGrid should not be engaged in (to avoid their efforts being spread too thin) and ways of leveraging work of partners and intermediaries rather than focusing only on end-users. The Campus Champions program was invoked in this regard, but was noted to have large holes in national coverage and no systematic organization, because it is more of a grass-roots enterprise. Note too that on-campus, small-scale computational resources are a necessity for jump-starting the use of high-end computing for science and engineering research, and all too many campuses do not have such facilities. Perhaps OCI should convene a broad group (including the Dept. of Education) to consider these issues around a more general cyberinfrastructure, including the role of cloud computing and state networks, etc. The national STEM need goes beyond what the TeraGrid can contemplate in isolation. This may be another point for collaboration with the ACCI task forces.

3. Future of NSF-Sponsored HPC

As noted at the TeraGrid SAB meeting in July 2009, there is considerable concern regarding the availability of sufficient HPC resources including computer cycles and data storage needed to underpin future NSF-sponsored HPC. In particular, the SAB is concerned about the possibility that the steadily growing demand for resources will substantially outstrip the future diminishing supply. There is also a serious concern about the availability of diverse computing platforms, e.g., symmetric multiprocessor systems. To frame the discussion of this issue, several presentations were made.

1. Kent Milfeld (Texas Advanced Computing Center, TACC) described the demand for TeraGrid resources by providing an overview of recent experience in the quarterly TeraGrid allocations. He pointed out that the past two allocations had seen less than half of the requested resources allocated, and he noted that the number of startup allocations had more than doubled in 2009 compared to 2007. Importantly, Milfeld showed how the trend in allocations has resulted in a significant increase in management and review activity, with over 500 TRAC allocation requests annually and collateral effects such as the need to adjust available resources to manage queue wait times.

2. Mike Norman (San Diego Supercomputer Center, SDSC) gave a presentation on the system called Gordon that is being built at San Diego Supercomputer Center and will be available in summer 2011. The system is intended to meet the needs for data-intensive tasks (data analysis workflows, data mining, visual analytics, on-demand data-driven applications, etc.) using an innovative design based on flash memory. The architecture includes 32 supernodes, each having 32 compute nodes (Intel Sandybridge) and providing an aggregate 8 TFLOPS capability accessing 2 TB random access memory (RAM) and 8 TB flash memory (solid state device, SSD). A prototype system called Dash will be used to assist the design and construction stages of Gordon.
3. Jeff Vetter (Georgia Tech Univ.) described the Keeneland system planned for installation at the National Institute for Computational Sciences (NICS) at the University of Tennessee-Knoxville – in partnership with Georgia Tech University, Oak Ridge National Laboratory (ORNL), nVidia and HP – in spring 2010 with the full-scale system (three times the size of the initial system) to be installed in spring 2012. The Keeneland system explores the feasibility of using graphical processing units (GPU – Fermi chip from nVidia), which provide a high floating-point result rate per unit power consumption (FLOP/watt) with high memory bandwidth. The project will focus on programmability, precision, accuracy and reliability. The initial system will be built with Intel Nehalem chips and nVidia servers with Fermi GPUs, connected via Infiniband 4X QDR. Each Fermi GPU has 512 CUDA cores and is expected to deliver more than eight times the double precision floating-point result rate of the current nVidia generation.
4. Craig Stewart (University of Indiana) spoke about FutureGrid, a grid computing experiment “factory” that he likened to the Shmoo (http://www.deniskitchen.com/docs/new_shmoofacts.html), created by Al Capp, which could be whatever its owner wants it to be. Stewart noted that FutureGrid is a research system and not intended as a production environment.
5. John Towns (National Center for Supercomputer Applications, NCSA) described the projected gap between available resources (supply) and allocation requests (demand) that can be expected based on trends in usage, and requests and actual allocatable systems that are available or expected to be deployed. He noted that the gap was measured as a factor of two (2X) in the latter half of 2009 when the old two-mode allocation system (LRAC and MRAC) was merged to a single quarterly TRAC process. Using best estimates and educated projections, Towns expects the gap to grow wider than 2X during 2010-2011. He considered the possibility of moving $\frac{1}{4}$ to $\frac{1}{2}$ of the future demand to campus-based systems, but he expected issues having to do with economies of scale and lack of willingness to buy systems (most campus chief information officers are only willing to host systems, not buy them) to make such offloading of demand challenging. Rob Pennington (NSF Office of Cyberinfrastructure, OCI) pointed out that, while about $\frac{2}{3}$ of the OCI budget went to HPC in 2007, less than $\frac{1}{2}$ of the 2010 budget will go to HPC.

The concern expressed in July 2009 remains valid, and is, in fact, more pressing with the decision by OCI not to make the Track 2C award to PSC due to a failure to achieve an agreement with the vendor on terms and conditions (www.nsf.gov/nsb/meetings/2009/0923/minutes.pdf). The OCI is gathering information from the community on how to address the Track 2C problem, while operating within the constraints of National Science Board (NSB) approval for awards greater than \$3 million and the overall budget that must meet a wide variety of demands beyond

computing hardware. It was noted that the ACCI has formed a task force on HPC that is gathering information on how to address the increasing gap.

It was noted that the TeraGrid is operated as a completely open resource, not limited to NSF-funded researchers. Strong opinions were expressed about the pros and cons of open vs. closed (NSF-only) access to the TeraGrid. On the one hand, the open access is very good for the NSF, because it enhances productivity and enables synergy within the communities being served. It also provides an opportunity for the NSF (Director) to seek a more open and jointly managed process for meeting the HPC needs of the U.S. research community. On the other hand, the openness to researchers not funded by NSF has attracted considerable resource demand from those communities that are funded by the National Institutes of Health, the Department of Energy, and other agencies, which have contributed considerably to the demand-supply gap. While it is difficult to determine the exact level of non-NSF usage of TeraGrid resources, using statistics from the most recent three TRAC allocation periods, approximately 30-40% of the NUs requested were associated with accomplishing the science objectives in grants from non-NSF sources of funding. In light of the current over-subscription of TeraGrid resources, and noting that other agencies place restrictions on computing resources they provide, the NSF might be forced to consider giving NSF-supported projects some degree of priority in the allocation process.

Based on the discussion, the SAB made the following statement:

“The NSF, via the open science TeraGrid and other, more discipline-oriented high-performance computing (HPC) facilities, supports fundamental computational science, based on merit review. The openness of the TeraGrid has been very successful, supporting computational research funded not only by the NSF, but also by several other federal and state agencies and, to a small degree, private companies. The TeraGrid Science Advisory Board (SAB) strongly supports this model of open access and strongly endorses continuing support by NSF for computational science research, both funded by NSF and funded by other agencies or entities. However, the SAB notes that the success of the open computational science model going forward depends on adequate resources being deployed for HPC facilities, including HPC systems and the software, networking and human expertise infrastructure that support them, which may be beyond the means of the NSF through its Office of Cyberinfrastructure (OCI). A gap already exists between the available HPC resources in the TeraGrid portfolio and the resources demanded through the allocation process. This gap will be exacerbated by the retirement of several smaller systems in the near future, and the recent failure of the Track 2C award, contract and procurement.

Accordingly, the SAB encourages the TeraGrid to

- further develop a mitigation plan that (a) informs users of the existing and planned resources and the anticipated gaps, (b) provides a clear timeline of available resources, in terms of allocatable computational power, and their expected lifetime, and (c) helps users identify alternative resources (campuses, agencies) that may be suitable for their research.

The SAB encourages OCI to

- act quickly to remedy the immediate problem associated with the Track 2C failure;
- work with the Cyberinfrastructure Coordinating Committee (CICC) of the NSF, and program directors in the relevant disciplinary divisions, to ensure that adequate computing resources are available and delivered to meritorious NSF-funded projects;

- work with the NSF Advisory Committee for Cyberinfrastructure (ACCI) in its task-force based strategic planning exercise to develop a long-term plan for delivery of high-quality, well-supported HPC resources; and
- reach out to other federal agencies, through the Office of Science and Technology Policy, or other interagency mechanisms, to develop a coordinated approach to support the Nation's growing requirements for fundamental computational science.”

4. Long-Term Data Storage

At its July 2009 meeting, the SAB expressed concern about long-term data storage and suggested introducing data management policies to provide incentives for use of existing persistent data storage and to encourage users to think carefully about the value of their data. This is a long-standing issue for the TeraGrid, however, since the TeraGrid has been funded primarily to provide HPC resources, so that long-term data storage has received relatively lower priority. It is clear that many open science researchers require data storage beyond the lifetime of computing allocations. However it is not clear whether TeraGrid RPs are required to provide such storage as part of their TeraGrid operational requirements.

To help frame the discussion, Chris Jordan (TACC) gave a short presentation on plans for archival replication service. The plan calls for a minimum of two copies of data sets at separate sites. The intent is to try a pilot implementation with 0.5-1.0 PB allocations for one year at 4-6 TeraGrid sites, which will be used to evaluate demand and resource constraints. The plan is being developed in collaboration with DataNet awardees to ensure that metadata standards and compatibility are addressed. Since the time of the SAB meeting, the replication service plan has evolved to include significant persistent disk resources, making it possible to address issues of access – by research communities and/or the general public – as well as preservation. Nonetheless, a dedicated planning effort is still needed to resolve issues such as the selection of access services to be provided and how to incorporate access concerns into the overall data distribution and replication scheme.

To further frame the discussion, a presentation by Phil Andrews (NICS) was given by Towns on a proposal for long-term archival storage, which was submitted to but not funded by NSF. The idea in the proposal was to establish a distributed virtual archive with multiple copies across the TeraGrid, thereby eliminating any dependence on a single resource provider. The virtual archive is intended as a critical resource to address the need that many users have to maintain persistent access to large data sets for 3-5 years after their creation. Several cost models could be employed, including passing on media costs to users.

In the discussion that ensued, a number of points were raised:

- This requirement is within the purview of the ACCI data task force.
- Many data sets of value to particular research communities have been created – whose responsibility is it to maintain, archive, replicate and disseminate such data sets? For example, the atmospheric sciences division of NSF supports the National Center for Atmospheric Research (NCAR), one mission of which is to support long-term archives of precious observations of the global atmosphere and world oceans.
- Often, the value in data sets is in their aggregation, which raises issues such as proprietary data rights, privacy, etc.

- Data sets are being generated in an ever-widening range of formats, e.g., GenBank is a database with strings of letters, which creates challenges for collection, archival, curation, etc.
- Similarly, different communities have different requirements and levels of readiness for addressing long-term data storage issues. Long-term storage can benefit some disciplines much more than others. This diversity argues in favor of a distributed approach.
- Data storage is a different commodity compared to computational cycles – the analogy made was that computing is like electricity (use it or lose it), while data storage is like cemetery plots. Nevertheless, it is becoming a critical resource, which means that it has to be allocated.
- Many of the challenges of long-term data storage have been addressed in the private sector – can this issue be out-sourced or can lessons be learned from commercial entities?
- Some data may be needed beyond the lifetime of the resource provider(s), where the long-term storage is provided.
- As data sets grow in size, the other factors in addition to storage – processing power, network bandwidth, etc. – must grow commensurately. “It takes a supercomputer to analyze the output of a supercomputer.” The TeraGrid is somewhat uniquely positioned in this regard, because it is a combination of HPC resources and high-speed networks. There also is a software aspect, e.g., complexity hiding for data dissemination and re-use.

The SAB recommends that the TeraGrid go forward with plans to prototype long-term data storage systems, maintaining close coordination with the ACCI task force on data. The TeraGrid should also consider alternative organizational models and arrangements, including partnerships with existing data archives and service centers. The SAB noted that there is a long-standing empirical relationship between the computational capability that is available to researchers and the volume of data added to the long-term archive, which argues in favor of a balanced approach when attempting to assign priorities among HPC, data storage, software and broadening participation.

The SAB recognizes that data tends to fill available storage, and this seems to be an inescapable, entropy-based principle. This suggests that regardless of the amount of storage purchased by TeraGrid RPs, there always will be value in encouraging efficient data management of existing data storage. The SAB recommends that as they deploy additional data storage, the TeraGrid and XD RPs introduce and strengthen efficient data management by users. The SAB feels that currently available storage might be more efficiently used by providing users with easy-to-understand reports and summaries on individual, and project-wide, storage usage. Policies might be introduced that change the default storage term from “indefinite” to “less than one year” to encourage users to remove files with low value. The TeraGrid might introduce a policy that requires more metadata for longer-term storage. The TeraGrid might also help users track information about data in storage such as “which files have backups”, and “cost in SUs to re-create a data set.” Introduction of policies like these, even implemented as user advisories, could help maximize use of any existing data storage.

5. Sustaining a Persistent National Cyberinfrastructure

To address a growing concern that the eXtreme Digital (XD) competition might be having an adverse impact on current TeraGrid operations, the SAB engaged in a discussion of how best to sustain a persistent national cyberinfrastructure. The discussion was framed by a presentation by Jay Boisseau (TACC) who reported on the recent workshop held by the ACCI task force on HPC. The outcome of the workshop was still in draft form at the time of the SAB meeting, but several points emerged that were worth noting. According to Boisseau:

- Scientific advances are well-served by competitive awards and peer-reviewed publications.
- On the other hand, sustaining scientific advancement through persistent growth in capabilities, innovation and impact requires a long-term strategy that may not be best accomplished by serial 3- or 5-year grant awards.
- The national HPC cyberinfrastructure is an enabler of open, diverse, transformative scientific research, not an end goal in and of itself.
- HPC cyberinfrastructure centers
 - attract and develop a cadre of rare expertise and blend it into effective teams, and
 - acquire and maintain long-term physical infrastructure of increasing scale and value.
- Therefore, there is a need to balance the competition for meritorious, new or re-designed HPC cyberinfrastructure centers and their distributed virtual organization against the value to transformative science and engineering, of having the experienced, advanced expert support teams and sustained infrastructure, albeit continually refreshed, that the science and engineering communities can include in their research and education plans and proposals.
- As pointed out by Sid Karin (founding director of the SDSC), there are no NSF supercomputer centers, there are only NSF-assisted supercomputer centers,
 - Pro: great flexibility and agility
 - Con: does not align well with very large awards of relatively short duration, heavy management, or academic institutions
- The workshop consensus was that cyberinfrastructure needs to be persistently supported, producing persistent growth in capabilities and maintaining user confidence in stability, which requires a long-term strategy with longer awards, a stronger focus on long-term data stewardship, a relatively small number of centers at the high end to take advantage of economies, capabilities and expertise concentration of scale.

In the discussion that ensued, several important points emerged:

- The requirements for co-located (vs. distributed) expertise need to be enumerated.
- Evolving user requirements must be addressed within the lifetime of longer center awards.
- As a distributed virtual organization, the TeraGrid is one of the most complex; therefore, the problems that the XD competition is trying to solve need to be articulated.

6. Identifying, Building and Sustaining Strategic Partnerships

The SAB had expressed an interest in the relationship between the TeraGrid and the Open Science Grid (OSG). Towns and Daniel Katz (University of Chicago) provided an update on the status of discussions, saying that the two groups are working to identify areas of collaboration and areas where the two groups are complementary. A white paper on workforce development, with a focus on students, is being written, and a joint proposal, driven by use cases, was submitted. The TeraGrid asked for guidance from the SAB regarding the nascent direction discussions with OSG are taking. The SAB expressed cautious support for the joint proposal idea, noting that moving slowly is the right way to go, because developing a relationship can take a lot of time. The SAB recommends that the TeraGrid take pains to communicate to users the benefits of a partnership with OSG, for example, enabling new activities, sharing computational load (moving high-throughput work from TeraGrid to OSG), and learning how to allocate resources more effectively (OSG has no allocation process at present). The SAB also suggested looking farther afield for organizations with which collaborations might be more valuable, although it was noted that since it is late in the TeraGrid funding cycle (also true for the OSG), so that new initiatives may have to wait for the XD award to be made.

7. Scientific Impact of the TeraGrid

For several meetings, the SAB has been discussing ways of documenting, measuring and improving the scientific impact of the TeraGrid. This is especially important in the allocation process. Towns explained that scientific impact is an elusive quantity for many reasons: it is more statistical in nature than traditional metrics like usage, allocations, etc.; the direct effect of an infrastructure on its users is conflated with other factors; the disciplines served by the TeraGrid are diverse and they measure impact in their own fields in different ways; and it has a longer time scale, requiring a long look back to determine impact.

The SAB suggested:

- Recruiting highly-qualified reviewers for the allocation process analogous to having highly-qualified editors for high-impact journals.
 - However, the SAB recognizes the difficulty in getting highly-qualified scientists to review infrastructure usage requests
- Compiling “gems” or “highlights,” that is, exemplary papers from various disciplines. For example, the TeraGrid could include on its web site links provided by its users to papers (to the extent that such links do not violate copyrights) that acknowledge or have benefitted from TeraGrid support.
- This list and or the list from the TeraGrid annual report could be ranked, possibly by the SAB or TRAC panelists.
 - A group of experts (e.g., SAB members) could write periodic opinions on papers worth reading that were enabled by TeraGrid resources
- Continue its “Science Highlights” publication of science and engineering brief case reports useful for those not in the particular field, the general populace, students, administrators, and federal and state legislators
- Adopting additional metrics, e.g., citation index, impact index of papers acknowledging TeraGrid support, success of grant proposals citing prior TeraGrid usage

- Having TeraGrid publish its own journal. Because the TeraGrid provides HPC resources and services to a wide range of computational science disciplines, there is an opportunity to provide a rigorous, academic venue where findings and lessons learned by TeraGrid users can be published. This goes beyond the possible venues published by Association of Computing Machinery (ACM) or IEEE, which are more focused on the computer science aspects than the computational science aspects. As an example of what can be gained, the Journal of Advances in Modeling Earth Systems (JAMES; <http://www.Adv-model-earth-syst.org>) is a new open-access journal focused on Earth system modeling and methods, which was started by an NSF Science and Technology Center as part of its required outreach activity – it was not that hard to do and could be duplicated as part of TeraGrid’s outreach.

8. Process for Selection of New Members of SAB

The current members of the TeraGrid SAB all have terms that will formally expire after the July 2010 meeting. The TeraGrid Forum has been working on a list of potential new members since summer 2009, seeking balance among disciplines, inclusion of people from under-represented groups, computational experience etc. A short list has been devised, but no invitations have been sent. The SAB recommends that the process of selecting SAB members:

- Include input from the NSF Cyberinfrastructure Coordinating Committee (CICC);
- Include representatives of other agencies, especially those that fund TeraGrid users or, at least, invite them to attend SAB meetings;
- Consider increasing the size of the SAB to increase participation at each meeting;
- Establish a set of criteria for population of the SAB that can be drawn from the following:
 - Knowledge about computing, familiarity with TeraGrid, but not just top 10 users of TeraGrid
 - Balance in gender and geographic distribution, especially EPSCoR states
 - Balance in disciplines, including those that aren’t using “big iron” today
 - Representation from industry providing HPC, using HPC or offering complementary services (e.g. cloud computing)
 - Representation from underrepresented minorities, persons with disabilities or the institutions that serve them
 - Representatives from academic computing centers
 - Senior/experienced people who have looked at work being done in community can serve as “representatives” of their discipline communities
 - Junior “rising stars” who are just starting out as HPC users
 - Experience with other computing environments, not just with TeraGrid
 - Representatives from other large programs such as PetaApps, SDCI, CDI, SciDAC or INCITE, or experience with such programs outside of the U.S.

Appendices

A.1 TeraGrid Science Advisory Board charge

The TeraGrid Science Advisory Board (SAB) is charged with:

- Providing 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;
- Considering the progress and quality of these activities, their balance, and the TeraGrid's interactions with the national and international research community;
- Providing advice on future TeraGrid plans;
- Identifying synergies between TeraGrid activities and related efforts in other agencies;
- Promoting the TeraGrid mission and its activities in the national and international community; and
- Providing help in nurturing and expanding the TeraGrid community.

A.2 TeraGrid Science Advisory Board membership

Eric Chassignet, Florida State Univ.

Thomas Cheatham Univ. of Utah

Gwen Jacobs, Montana State Univ.

Dave Kaeli, Northeastern Univ.

Jim Kinter (chair), Center for Ocean-Land-Atmosphere Studies & George Mason Univ.

Luis Lehner, Louisiana State Univ.

Michael Macy, Cornell Univ.

Phil Maechling, Univ. of Southern California

Alex Ramirez, Hispanic Assoc. of Colleges and Universities

Nora Sabelli, SRI International

Pat Teller, Univ. of Texas, El Paso

P. K. Yeung, Georgia Institute of Technology

Cathy Wu, Univ. of Delaware and Georgetown Univ.

A.3 Meeting agenda

Thursday, 10 December 2009

8:30 am CONVENE; introductions

8:45 am Broadening participation, specifically in TeraGrid, and more generally in computational science and high-performance computing, by

- a. providing access to new communities, often understood as "underserved" in terms of science and technology;
- b. educating the next generation of scientists and technologists; and
- c. integrating the TeraGrid and campus-based computing infrastructures.

Questions for the SAB:

- i. What can/should TeraGrid do now? In the future?
- ii. What is the likely impact on current users of a broader user community?
- iii. How does broadening participation impact, if at all, allocation policies w.r.t. NSF- and non-NSF-sponsored projects?

10:00 am BREAK

10:30 am Resume discussion of broadening participation

11:30 am Future of NSF-sponsored HPC on the ...

- a. Demand side – report on recent resource requests and allocations; possible survey of user requirements
- b. Supply side – report on plans for new resources to be deployed (SUs; platform diversity)

Questions for the SAB:

- i. What are the gaps in current TG portfolio of HPC resources?
- ii. Where should NSF apply future funding to best meet the needs of the science and engineering community?
- iii. What should NSF make sure is provided by XD?
- iv. What do users like or dislike about the current TG (resources, management) and future TG/XD (e.g. ease of transition to new architectures)?

12:30 pm	LUNCH
1:30 pm	Resume discussion of future of NSF-sponsored HPC
2:30 pm	Long-term data storage <ul style="list-style-type: none"> a. Report on data management policies to provide incentives for use of existing persistent data storage and to encourage users to think carefully about the value of their data b. Report on plans for new resources
	<i>Questions for the SAB:</i> <ul style="list-style-type: none"> i. Is there a requirement for data storage beyond the lifetime of computing allocations? ii. If so, how should it be addressed?
3:30 pm	BREAK
4:00 pm	Resume discussion of long-term data storage
5:00 pm	ADJOURN
6:30 pm	Dinner (Pinzimini restaurant)

Friday, 11 December 2009

8:30 am	CONVENE
8:45 am	Impact of eXtreme Digital (XD) competition on effective operation of the TeraGrid <p><i>Questions for the SAB:</i></p> <ul style="list-style-type: none"> i. Is there a discernible negative impact of this competition on current operation? ii. If so, how should it be addressed?
10:00 am	BREAK
10:30 am	Identifying, building and sustaining strategic partnerships, especially with other national and international cyberinfrastructure networks <ul style="list-style-type: none"> a. Status report on progress in TG/OSG relations, including report from OSG Joint Oversight Team meeting (11/5/2009) b. Broader formal and informal partnering opportunities
11:30 am	Scientific impact of the TeraGrid – report on how to measure/document scientific impact of the use of TeraGrid resources
12:30 pm	ADJOURN