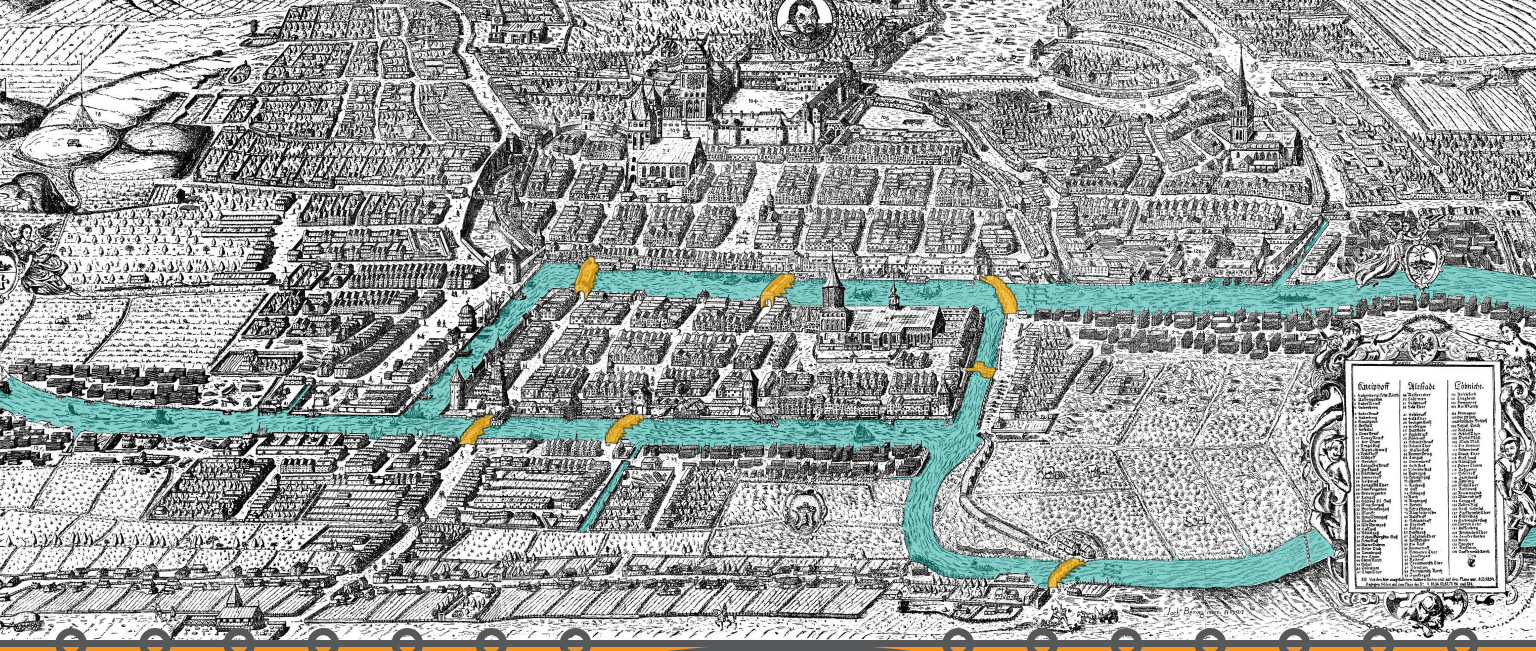


Campus Bridging

Software & Software Service Issues Workshop Report

August 26-27, 2010
Denver, Colorado

Editors: John McGee, Von Welch, and Guy Almes

Please cite as:

McGee, J., Welch, V., Almes, G. (eds). Campus Bridging: Software & Software Service Issues Workshop Report. 2011. Available from: http://hdl.handle.net/2022/13070

Acknowledgments

The workshop organizers would like to thank Dale Lantrip for handling logistics for the workshop, Ray Sheppard and Jonathan Morrison for taking notes, Bill Barnett for helping to moderate the workshop, and Nancy Bannister of the Indiana University Survey Research Center for her efforts with the Cyberinfrastructure (CI) user survey.

The workshop organizers would like to thank all the NSF researchers who took the time to respond to the CI survey.

Finally, we thank Malinda Lingwall, Richard Knepper and Peg Lindenlaub of the Indiana University Pervasive Technology Institute and Maria Morris of Indiana University Creative Services for their support in the production of the final report document.

The preparation of this report and related documents was supported by several sources, including:

- The National Science Foundation through Grant 0829462 (Bradley C. Wheeler, PI; Geoffrey Brown, Craig A. Stewart, Beth Plale, Dennis Gannon Co-PIs).
- Indiana University Pervasive Technology Institute (http://pti.iu.edu/) for funding staff providing logistical support of the task force activities, writing and editorial staff, and layout and production the final report document.
- RENCI (the Renaissance Computing Institute, http://www.renci.org/) supported this workshop and report by generously providing the time and effort of John McGee.
- Texas A&M University (http://www.tamu.edu) supported this workshop and report by generously providing the time and effort of Guy Almes.

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, the Indiana University Pervasive Technology Institute, or Indiana University.

Other materials related to campus bridging may be found at: https://pti.iu.edu/campusbridging/

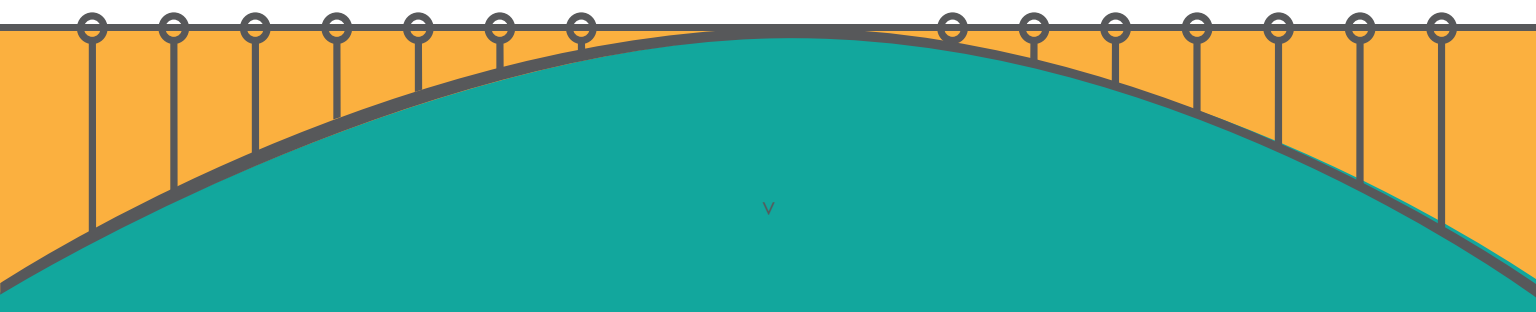
Workshop participants

The 32 workshop participants and their affiliations are listed below.

Guy Almes, Texas A&M U.
William Barnett, Indiana U.
David Chaffin, U. of Arkansas
Tom Cheatham, U. of Utah
Gary Crane, SURA
Kim Dillman, Purdue U.
Dan Fraser, Open Science Grid
Brandon George, U. of Oklahoma
Andrew Grimshaw, U. of Virginia
Gary Jung, LBNL
Rajendar Kanukanti, LONI
Daniel S. Katz, U. of Chicago/TeraGrid
Ken Klingenstein, Internet2
Bill Labate, UCLA
Dale Lantrip, Indiana U.
Cliff Lynch, CNI
Miron Livny, U. of Wisconsin
Man Luo, U. of North Carolina
John McGee, RENCI
Amit Majumdar, SDSC/TeraGrid
Jonathan Morrison, Indiana U.
Michael Mundrane, UC-Berkeley
Jan Odegard, Rice University
Vani Panguluri, Southern U.
Jim Pepin, Clemson University
John-Paul Robinson, U. of Alabama-Birmingham/SURA
Traci Ruthkoski, U. of Michigan
Ray Sheppard, Indiana U.
Steve Simms, Indiana U.
Martin Swany, U. of Delaware
David Walker, UC-Davis
Von Welch, Independent

Organizing committee

The workshop organizing committee consisted of John McGee (RENCI), Craig Stewart (Indiana U.), and Von Welch (independent consultant). Bill Barnett (Indiana U.) helped moderate the workshop. Dale Lantrip (Indiana U.) handled logistics for the workshop.



Executive Summary

The NSF Advisory Committee on Cyberinfrastructure (CI), as part of its process to create a Cyberinfrastructure Framework for 21st Century Science and Engineering, established six task forces to investigate various aspects of the application of cyberinfrastructure to enable scientific discovery. One of these task forces is the Task Force on Campus Bridging. That task force [1] has defined campus bridging in the following way:

> *The goal of campus bridging is to enable the seamlessly integrated use among: a scientist or engineer's personal cyberinfrastructure; cyberinfrastructure on the scientist's campus; cyberinfrastructure at other campuses; and cyberinfrastructure at the regional, national, and international levels; so that they all function as if they were proximate to the scientist. When working within the context of a Virtual Organization (VO), the goal of campus bridging is to make the 'virtual' aspect of the organization irrelevant (or helpful) to the work of the VO.*

This initiative by the NSF has led to several new workshops and new research in cyberinfrastructure. This report summarizes the discussion at and findings of a workshop on the software and services aspects of cyberinfrastructure as they apply to campus bridging. The workshop took a broad view of software and services, including services in the business sense of the word, such as user support, in addition to information technology services. Specifically, the workshop addressed the following two goals:

- Suggest common elements of software stacks widely usable across the nation/world to promote interoperability/economy of scale; and
- Suggested policy documents that any research university should have in place.

To guide the workshop discussions, the workshop organizers used an online user survey designed to capture user experiences with CI and sent to 5,000 scientists who have served as NSF PIs. The survey identified how CI is used to bridge and which aspects of CI were working well and not as well. Nearly half of the respondents indicated they used some CI besides their own workstation or locally controlled CI. An analysis is included in this report with details given in Appendix 1.

Resulting from this survey and workshop discussions, a number of findings emerged, which we group into the following four categories:

Challenges related to software and services for campus bridging. (1) Scientists have no coordinated mechanism to discover CI resources and services, and, once discovered, it is a challenge to figure out how to use those resources, determine their policies, find users support, etc. (2) Conversely, it is difficult for CI projects to discover small communities and discern the needs of those communities (as opposed to their vocal minorities). (3) There are significant challenges in measuring the effort spent on and impact of campus bridging and campus-level CI due to the distributed nature of these activities and lack of clear metrics. (4) Scientists are hampered in their use of CI by a lack of coordination and interoperability between CI and campus support mechanisms.

Importance of campuses in CI education and workforce development. (1) People, as constituents of a trained workforce, are critical to mature, usable CI. (2) The new generation of scientists and students are accustomed to commercial and other computing infrastructure that raises expectations on CI. These scientists also tend to be less accustomed to the low-level usage modalities used by their predecessors. (3) Campuses, as educators of the future CI workforce and scientists, can have huge impacts on both of the previous findings.

Relationship of NSF CI, administrative computing, and commercial CI. (1) Computing infrastructure within campuses tends to be split between administrative and research computing. Research computing is more readily integrated as part of a coordinated national CI, but the administrative IT tends to get more attention and funding from campus leadership. (2) Administrative and commercial computing infrastructure tend to have strong non-functional attributes (reliability, usability, etc.). This has led to significant adoption of commercial infrastructure by scientists as indicated in the CI user survey. (3) While the non-functional attributes of CI should be improved to meet scientist's increasing expectations, CI needs to maintain enough flexibility to adapt to the still-evolving needs of collaborative science.

User support and fostering expertise sharing for CI. Effective use of CI for science requires good user support and access to CI expertise. The reward system for many faculty PIs, however, does not motivate supporting scientists. Changing this reward system would be very difficult. Other ways to increase support for scientists using CI would be to provide for more "peer-to-peer" support through user forums and the like, and increasing the CI expertise of campus support staff close to scientists and giving that support staff cognizance of the scientist's problems with CI outside the campus.

Emerging from these findings were the following recommendations:

- *NSF must lead the establishment of a coordinated, national cyberinfrastructure support system to provide user support and expert guidance for using cyberinfrastructure.* This system should be constructed in concert with and with contributions from campuses, regional providers, and CI projects. This support structure must be constructed with the backing of campus leadership and coordinated with campus support services to bring support as close to scientists as possible. The system must be neutral with respect to any given CI project or technology in order to put the need of the scientists first.
- *NSF must lead the community in establishing a blueprint for a National Cyberinfrastructure.* A "National Cyberinfrastructure" composed of CI from state, regional, and federal sources must be coordinated to be effective. Campuses and other CI providers need a better understanding of what the National CI is in order to effectively integrate their activities. A blueprint would provide architecture for this National CI, showing how the different contributions contribute. This blueprint must be trusted to be neutral to any particular project or technology viewpoint, and focused on furthering domain science rather than CS research. There are significant

challenges in selecting an appropriate body to create this blueprint to meet these goals as CI expertise tends to be integrated with CI projects and hence seen as biased.

- *NSF must continue to emphasize maturity (reliability, usability, etc.) in its review process for cyberinfrastructure.* The new generation of scientists and students are accustomed to more mature commercial and administrative campus computing infrastructure. CI needs to increase its level of robustness towards this new bar, while maintaining enough flexibility to adapt to evolving demands of collaborative science.
- *NSF must continue to provide leadership towards a national cyberinfrastructure.* NSF's technical vision was just as critical as its funding for the establishment of the NSFNET. While providing such leadership is significantly more challenging at this time, NSF's voice has impact beyond its funding programs and by continuing to provide a vision and guidance, NSF can significantly advance CI outside of direct funding efforts.

1. Introduction

As laid out in the National Science Foundation's (NSF) "Dear Colleague Letter: Cyberinfrastructure Vision for 21st Century Discovery," [2] cyberinfrastructure (CI) is a key and necessary component to support science and engineering. In the same document, NSF set for itself a vision to lead the development of a comprehensive cyberinfrastructure: "NSF will play a leadership role in the development and support of a comprehensive cyberinfrastructure essential to 21st century advances in science and engineering research and education." In support of this vision, the NSF Advisory Committee on Cyberinfrastructure (ACCI) created a set of six task forces to investigate various aspects of the development of cyberinfrastructure, including the Task Force on Campus Bridging.

The goal of campus bridging is to enable the seamlessly integrated use among: a scientist or engineer's personal cyberinfrastructure; cyberinfrastructure on the scientist's campus; cyberinfrastructure at other campuses; and cyberinfrastructure at the regional, national, and international levels; so that they all function as if they were proximate to the scientist. When working within the context of a Virtual Organization (VO), the goal of campus bridging is to make the 'virtual' aspect of the organization irrelevant (or helpful) to the work of the VO. Campus bridging is critical to supporting the ever-increasing level of cross-disciplinary and cross-organizational aspects of scientific research, as it enables not just the connection of scientists with CI beyond their campus, but also the connection of scientists with other scientists to support collaboration.

In August 2010, Indiana University coordinated a workshop on the software and services aspects of CI supporting intra- and inter-campus scientific collaboration as a followup to an earlier workshop on cyberinfrastructure software sustainability [3]. The workshop took a broad view of software and services – in particular using the term "services" to include the business sense of the word and encompassing services such as user support and Campus Champions, in addition to information technology services. Specifically, the workshop addressed the following two goals:

- Suggest common elements of software stacks widely usable across nation/world to promote interoperability/economy of scale; and
- Suggested policy documents that any research university should have in place.

In the remainder of this report, we present the survey of CI usage by NSF scientists conducted by the organizers as input to the workshop, and the workshop findings and recommendations. Workshop materials, including presentations and submitted white papers, may be found on the Campus Bridging web site [1].

2. CI user survey

To help guide the workshop discussions, the workshop organizers developed and deployed an online user survey designed to capture user experiences with CI, identifying how it is utilized to bridge and which aspects of CI were working well and not as well. An invitation to take the survey was sent to 5,000 NSF-funded scientists as well as being made available publicly on the workshop website, allowing anyone wishing to take it to do so. The set of 5,000 scientists was randomly selected from a pool of 34,623 people classified as principal investigators funded by the NSF between 1 January 2005 and 31 December 2009, and for whom we had a valid email address (excluding 1224 PIs for whom we could not determine a valid email address). The survey was implemented by the Indiana University Center for Survey Research under Indiana University Institutional Review Board approval.

The survey was initiated on 5 August 2010. Immediately prior to the workshop, a snapshot of 710 survey responses that had been collected thus far was used for presentation at the workshop. The survey continued after the workshop until 24 September 2010, when a total of 1,387 responses had been collected. We briefly discuss the survey results in this section; please see Appendix 1 for a complete review of the survey data including the questions regarding barriers to CI adoption, textual answers provided by the respondents, and a discussion on lessons learned from the survey process and how it could be improved in the future.

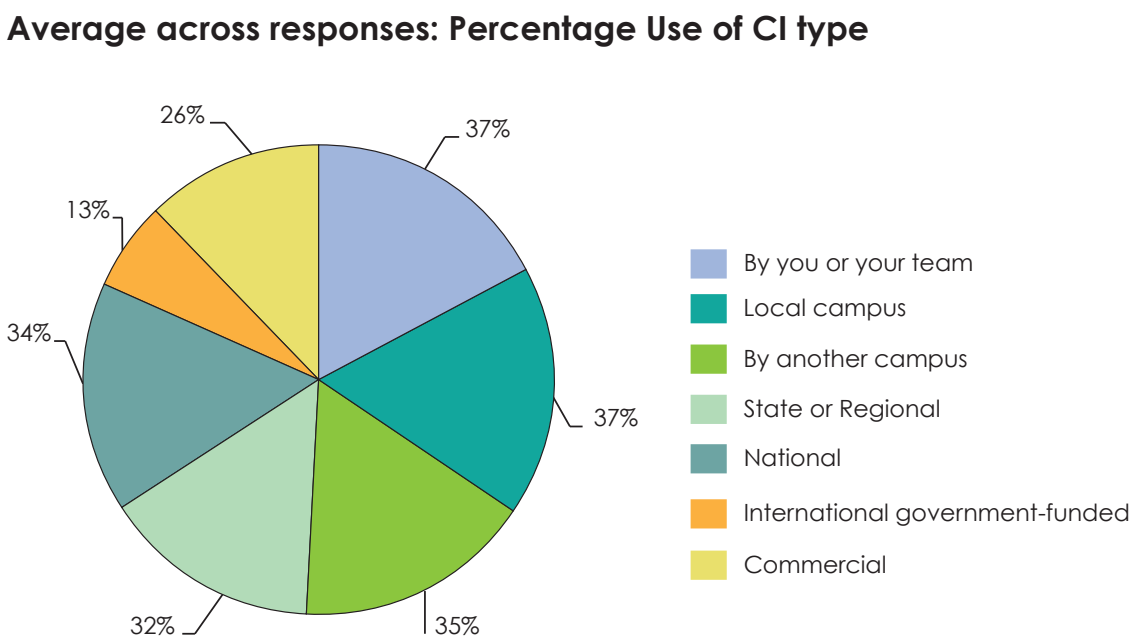


Figure 1. Percentages of respondents who indicated they used CI operated by different types of entities. The total is greater than 100%, indicating respondents who responded that they used CI operated by multiple types of entities.

Perhaps the most surprising result is that 722 (53%) of the respondents indicated that they do not use CI beyond their personal workstation or other CI operated by themselves and their team. We believe that at least some of these individuals who responded in this way did not consider the daily usage of networks, community datasets, communication/collaboration mechanisms, etc. as CI. A useful follow-up activity would be to solicit these respondents to take a new survey to better understand this issue.

Figure 1, showing type of CI usage cross-referenced with who operates the CI, indicates that overall CI usage is relatively evenly distributed across the "who operates the CI" dimension.

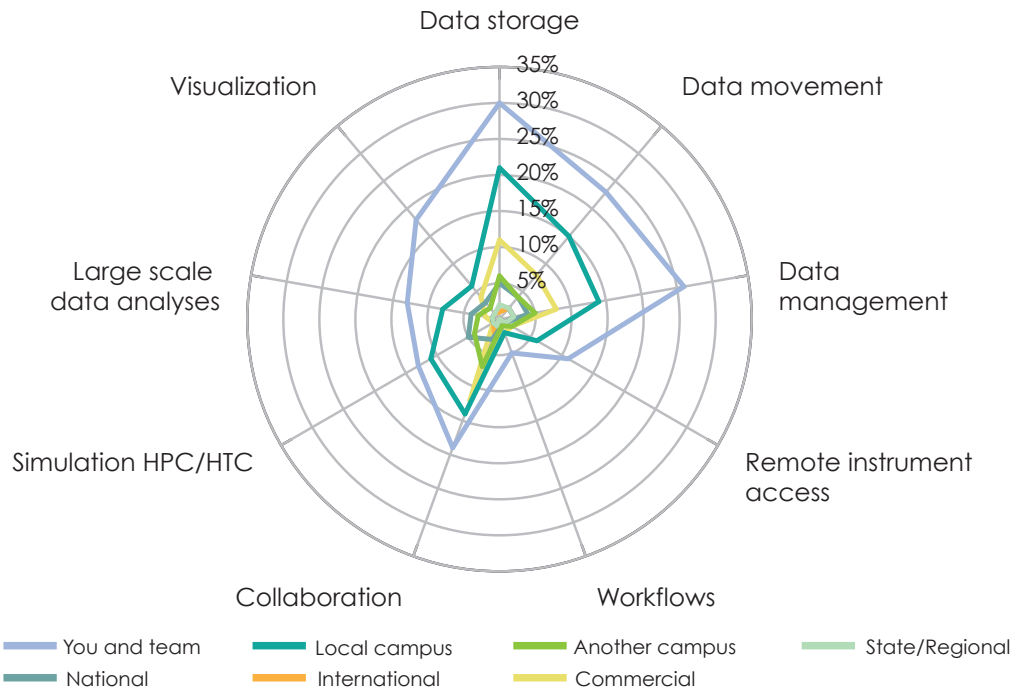


Figure 2. Type of CI cross-referenced by type of operator. The figures shows data storage, movement and management are highly concentrated local to the researchers.

Figure 2 indicates that data storage, movement, and management is highly concentrated at the individual, local team, and local campus level, with commercial providers coming in significantly less, yet still meaningfully more than "another campus" and national resources. In this data space, state/regional and international are nearly negligible. The commercial CI providers show a significant usage in collaboration tooling, and not surprisingly, the national resources are used most heavily for simulation, data storage, and large-scale data analysis. Outside of the local campus, access to remote instruments was most often accommodated by another campus.

As indicated in Figure 3, the most common method of accessing CI across the entire dimension of providers is via Web browser/portal. With the exception of commercially provided CI, the next most highly used access method across the providers is SSH.

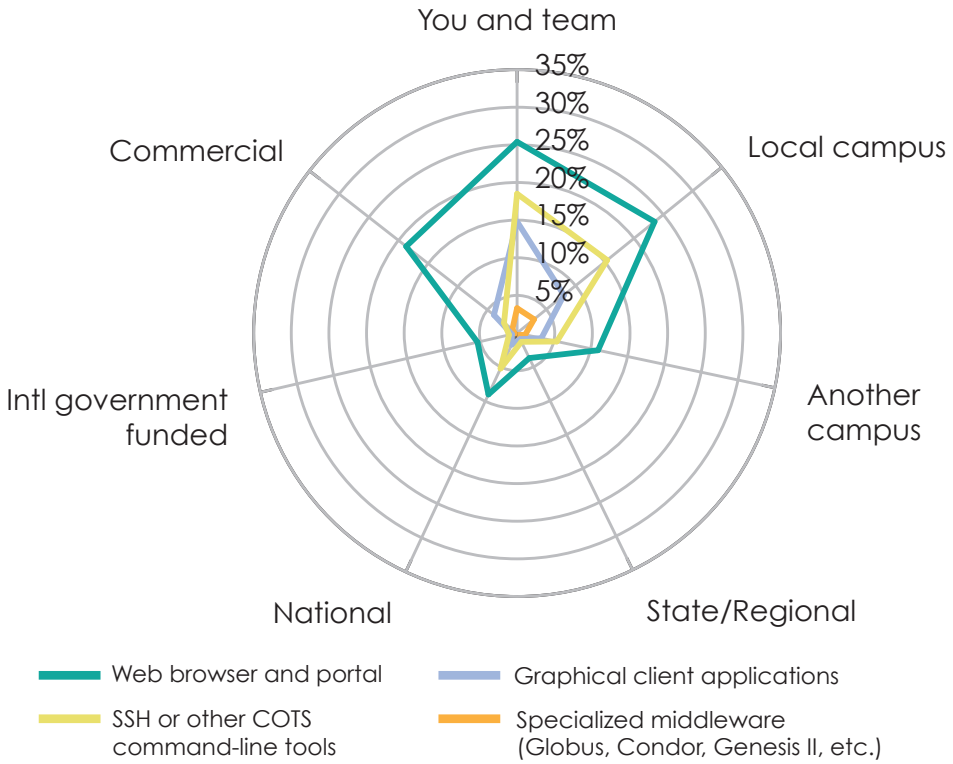


Figure 3. Access method cross-referenced by type of CI operator. The figure shows that Web-based access is most common, followed by SSH except in the commercial space.

3. Workshop findings

Here we present workshop findings organized into four major topic areas by subsection. The subsequent section gives the workshop recommendations that resulted from these findings and discussions.

3.1. Challenges related to software and services for campus bridging

Campus bridging is explicitly about CI supporting science and scientific collaboration that crosses domain boundaries among (and within) organizations. The crossing of boundaries is a common source of challenges. These often, for example, include network firewalls. These also, however, include difficulties in discovering expertise, services and other aspects of CI outside a scientist's local domain. Some specific challenges emerging from discussions for software and services to support science were:

- Scientists have difficulty locating resources and services. Each campus' support infrastructure reasonably tends to focus on the mechanisms they have chosen to deploy locally and there is no coordinated mechanism to locate available services on the national level (other campuses, regions, and federally-funded CI) that scientists and support staff can turn to for help.
- Even when scientists discover services, determining how to use them is a challenge. Mechanisms for describing interfaces, run-time environments and policies (e.g., access control, scheduling) for the resources are not standardized and, when they exist, are hard to locate. Also, CI support services and expertise to help the scientist with the use of those resources are difficult to discover. Adding to this challenge is that CI expertise is often specific to particular pieces of technology and projects, and may have conflicts of interest in advancing a particular approach as opposed to making neutral recommendations to the greatest benefit of the scientist.
- Conversely, it may be difficult for CI providers to find communities of users, either in the form of small organizations or science domain communities. Distinguishing between the needs of the broader community as opposed to those of a vocal minority is hard.
- Measuring effort spent on campus bridging and research computing and the resulting impact on science is challenging. The effort is distributed and much occurs in the form of students or research staff working alone or in small teams, whose time is not generally measured or collected. Also, there is no clear set of meaningful metrics for measuring the impact of the CI on science, and collecting information on those metrics is again difficult because of the distributed nature of the CI and science.
- Supporting science is hampered by a lack of interoperability and coordination among institutional and project support infrastructures. For example, there is no facility to allow support staff to find each other across domains (e.g., between a campus and a CI project) or enable the exchange of support tickets among systems to allow support staff at an institution to be aware that a local scientist is having difficulty and has opened a support ticket with a CI project.

- The reward system for faculty developing CI does not encourage effort spent on supporting users of the CI. Changing that rewards system would be difficult, and finding other avenues to support that CI should be explored.

3.2. Importance of campuses in CI education and workforce development

Discussions during the workshop emphasized the critical role of people in CI and the importance of their CI-related education and training. One aspect is that a well-trained workforce is critical to make CI a sustained infrastructure, with the stability, robustness and predictability the term infrastructure implies and scientists increasingly expect. The observation was made that hardware resources are increasingly cheap, but the lack of a qualified workforce for CI continues to be a challenge. Students are not generally exposed to the task of developing reliable software and systems. Their education focuses on research and prototyping; this has an impact "upstream" when these students later enter the workforce.

A second aspect of CI education is the level and type of familiarity that scientists have with regards to CI. For example, new students, who will comprise tomorrow's scientists, having grown up with a mature Internet and Web-based services, have both greater expectations of computational services in terms of usage modalities (e.g., command-line versus web), robustness, ease-of-use, etc. and, in general, less familiarity with complicated computer science concepts, command-line interfaces, and low-level development.

Campuses play a critical role in training and education, both in terms of developing that workforce for CI, and in training both CS and domain scientists in what CI can do for them and how to take advantage of it. A successful strategy for enabling campus bridging needs to be providing campuses with both the motivation and expertise to provide appropriate education with regards to CI; for example, the University of Virginia has a CS 101 course for graduate students across disciplines.

3.3. Relationship of NSF CI, administrative computing and commercial CI

Campus computing infrastructure can generally be categorized as administrative computing, supporting business functions such as email, payroll, and enrollment; and research computing, supporting scientific research. In general, research computing infrastructure is more readily usable as part of a coordinated CI than is administrative computing infrastructure (a notable exception is identity federation). Campus leadership generally puts more emphasis on and resources behind administrative computing than research computing. While this is certainly reasonable given the strong requirements for supporting the institution on administrative computing, the workshop participants agreed that the gap could be narrowed. Outreach to CIOs, VPRs, and other campus leadership on the benefits to science of research computing could help generate more balanced levels of support for research computing on campuses. CI, in turn, could leverage the increased

research computing. Additionally, methods for leveraging administrative computing (e.g., the way federated identity leverages campus identity management infrastructure) could be explored.

One observation was that both campus administrative computing infrastructure and commercial computing infrastructure have become very robust and have raised user expectations with regards to non-functional requirements (reliability, predictability, etc.). Commercial infrastructure's success in addressing these non-functional requirements was credited with its adoption, as demonstrated in the CI user survey. The workshop participants discussed the question of to what degree CI needs to match the increased maturity of administrative and commercial computing. It was unresolved whether closing this gap completely is the right solution. In favor of closing the gap is the observation that integrating administrative and research computing could bring some of the resources behind administrative computing to bear on making research computing and CI more robust, improving its non-functional attributes, all of which would encourage their adoption. It was pointed out, however, that administrative infrastructure and CI are optimized for different activities, with different risk strategies, and the conservative nature of administrative computing may not be appropriate for CI. There is a need to balance flexibility and advancement of CI (e.g., adoption of new technology and features) to support the dynamic nature of distributed scientific research with the risk to reliability, availability and robustness caused by changes. The conclusion was that the balance desired for CI is probably different than the balance desired for typical campus administrative computing, though quantification remains a challenge.

The workshop, however, generally agreed that today's CI, if it is to maintain and grow user trust and adoption, needs to demonstrate more of the non-functional attributes to meet the increased user expectations fostered by the more mature campus and commercial infrastructure. For a scientist, adopting CI as part of a workflow makes that CI critical to their science (e.g., papers they may publish, Nobel prizes they are competing for), which is a significant assumed risk, particularly with CI that is not under their direct control (e.g., CI operated by the campus, a regional or national provider).

3.4. User support and fostering expertise sharing for CI

Providing scientists with good user support and access to CI expertise for more advanced needs is critical for achieving successful science with CI. As discussed previously in Section 3.1, however, it is difficult for scientists to locate CI expertise and support where they are unfamiliar with projects and staff. Contributing to this is the difficulty of obtaining funding for user support and also a lack of interest in providing user support by many faculty PIs since it is outside their tenure reward system.

Generating more "peer-to-peer" support of scientists by fostering communities that can help each other is one strategy to provide more help to scientists. The challenge with this approach is that scientists who have developed expertise are already busy with their own science. Establishing mechanisms to encourage the formation of communities might help in this area.

Increasing CI expertise of local campus support staff is another strategy to improve support for scientist use of CI. Campus support staff, however, have the same challenge that the scientists do in finding external expertise, documentation and so forth on CI that is not used locally on campus. Campus staff would also benefit from increasing the visibility of issues among support staff, for example allowing local support staff to be aware when scientists on their campus contact CI project support staff.

4. Workshop recommendations

The goals of the workshop, reiterated from the introduction, were:

- Suggest common elements of software stacks widely usable across nation/world to promote interoperability/economy of scale; and
- Suggested policy documents that any research university should have in place.

In response to these goals, the workshop participants put forward the following recommendations, each captured in their own subsection.

4.1. The NSF should lead the establishment of a national CI support system

Currently user support for CI is very distributed and uncoordinated, presenting scientists with a substantial challenge in finding unbiased expertise when going outside their institution. Programs such as the U.K. Campus Grid Special Interest Group [4], the NIH Knowledge Centers [5], and the TeraGrid Campus Champion [6] programs have shown promise in addressing this problem but are limited to specific CI projects and technologies.

The workshop recommends that the NSF foster the establishment of a CI support system coordinated at the national level, comprised of expertise at campuses and CI projects, with sufficient vertical and horizontal coordination to allow for the routing of scientist's problems with CI to the appropriate expert with minimal effort by the scientist and support staff. Such a program should particularly seek to foster expertise on campuses, close to scientists, with buy-in from campus leadership to ensure that support staff have their role appropriately prioritized.

There are some unknowns in the exact methodology for such a service, but the participants agreed on some of its attributes:

- It is important that the support service be both technology and project neutral so that scientists trust the service to provide answers that best support their science rather than providing success stories for CI providers.
- A minimum investment to setting up such a service would be providing CI training (with travel expenses) and recognition to support staff. This would allow staff to develop both the technical expertise for answering user questions directly, along with the personal connections for finding the expertise for questions outside of their knowledge.
- This service should provide a feedback mechanism for gathering and aggregating experiences from the scientists using the CI in order to provide feedback for the larger CI ecosystem as to both successes, to help with impact assessment, and frustrations, for recognizing where improvement is needed.

4.2. The NSF should lead the establishment of a national CI blueprint

The move of campuses, at the insistence of the NSF, to adopt TCP/IP in the 1980s and, more recently, their move to adopt federated identity and InCommon are examples of campuses providing services that contributed to a National CI, comprised of coordinated CI at the campus, state, regional, and national levels, and pursuant to a national CI blueprint. While a number of different services (e.g., namespaces, databases, file systems, allocations policies) were discussed at the workshop, attempts to prioritize these services and their benefit to CI was stymied by a lack of agreement as to what constituted a national-scale CI. While participants had their own visions, there clearly wasn't agreement either among the participants or, it was agreed, among the broader community working on CI.

This lack of agreed-to vision for national CI has impacts that extend beyond the workshop participants: it means campuses have no single voice to follow to deploy both their own campus CI and contribute to a national infrastructure. Workshop participants agreed that for CI to succeed, there needs to be a driving vision, principles, use cases, and a blueprint that is accepted across the NSF directorates, the CI development community, and national campus leadership.

There was also agreement that there is no obvious body to develop such a blueprint today. Some of the challenges in selecting such a body and developing the blueprint that were discussed included:

- For the blueprint to be accepted, it should be generated in such a manner that people trust that it is neutral to any particular agenda and not designed to be overly complex for the sake of larger funding. It is difficult, however, to find expertise whose neutrality isn't compromised by being tied to a project or technology. Using a committee or standards body (e.g., Open Grid Forum) as a mechanism to develop a neutral blueprint is an option that has been attempted, but the weakness of this approach is that it tends to end either in deadlock or with a "design by committee" compromise.
- The blueprint needs to consider not only NSF science needs but also the CI of other federal agencies and their science drivers. History shows that such coordination is not without challenges.
- A blueprint needs to be sustained in order to adapt to evolving technologies, science needs, and deployments by campuses, commercial entities and other federal entities.
- CI providers have their own goals and deadlines that may compete with coordination. A blueprint will not be a panacea, but will enable choices and architecture to be coordinated when possible.

Despite these challenges, it was agreed that the development of a CI blueprint is critical to a national-scale CI and NSF clearly has a role in leading the community in its development.

4.3. The NSF should encourage mature CI in their review processes

In order to meet the rising expectations of scientists, CI needs to increase its maturity (non-functional attributes such as reliability and usability). Fostering this maturity is more than funding more software development; in fact, it may mean funding less development and putting more emphasis on non-functional attributes of the software that is developed. Some suggestions made to improve the maturity of CI produced under NSF funding included having a review process include an evaluation of whether the proposed CI is appropriately leveraging existing CI (e.g., does it integrate well with existing support services); increasing the evaluation on effectiveness of the proposed CI to support science rather than its speed, novelty or CS research impact; evaluating how the proposed CI contributes to the overall infrastructure; and increased emphasis on sustainability, possibly by finding methods to make campuses, instead of individual PIs, feel ownership of results; continue to encourage adoption of standard definitions and methods of describing resource runtime environments, usage policies, etc. to allow for easy migration between resources.

4.4. The NSF should continue to lead

Despite the fact that campuses have priorities driven by a number of other agencies and requirements, participants agreed that NSF's voice is still very effective in providing leadership to campuses, even when that leadership was just in the form of providing vision and guidance as opposed to fully funding initiatives. As with the adoption of TCP/IP, NSF should continue to show leadership in ways beyond funding, for example, by defining a CI blueprint as discussed in the previous recommendation. Coordination of that leadership with other agencies would magnify the strength of NSF's voice.

5. Conclusion

CI has made significant advances in enabling bridging between campuses to support increasingly important collaborative science research. For CI to continue to keep pace with maturing administrative and commercial computing, it needs to mature in its non-functional attributes (reliability, usability, etc.) and the human processes of user support, training, and education that support and enable its use. Leadership is also needed in the spirit of the early adoption of TCP/IP in the 1980s in order to coordinate community CI development and deployment.

Appendix 1. User CI survey

A survey of NSF scientists was conducted in preparation for the Campus Bridging Workshop on Software and Services to help inform the discussions. The survey was implemented by the Indiana University Center for Survey Research (IUCSR) under Indiana University Institutional Review Board approval. A set of 5,000 scientists were randomly selected from a pool of 34,623 people classified as principal investigators (PIs) funded by the NSF between 1 January 2005 and 31 December 2009, and for whom we had a valid email address (excluding 1224 PIs for whom we could not determine a valid email address). The survey was initiated on 5 August 2010 and participation was requested via e-mail. During the course of the survey three reminder messages were sent to the individuals who had not yet responded. A snapshot of the results was taken on 21 August 2010 for analysis and presentation at the workshop, while the survey continued to be active until full closure on 24 September 2010. The results presented here supersede the results presented at the 25 August 2010 workshop and represent the complete set of data collected during the term of the survey. As of the writing of this report, the survey continues to be available at the following location:

https://iucsr.qualtrics.com/SE/?SID=SV_6fAnne3gmd88f1G

Appendix 1.1. Survey results

The first question asked if the respondent uses any CI beyond their personal workstation or other CI operated by themselves and their team. If the respondent selected "no" to this question, they were directed to the end of the survey. Respondents who selected "yes" were guided through a series of questions for which results are shown below. 47% of the 1,387 total respondents selected "yes."

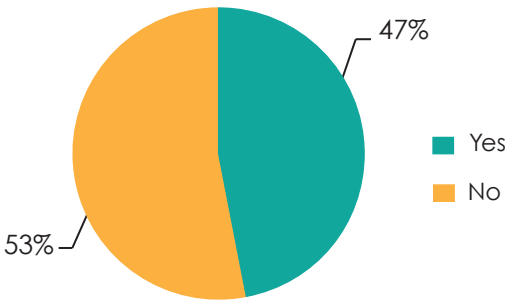


Figure A.1.1.

The 665 respondents who selected yes to question one were asked to specify the percentage of their CI usage as follows: operated by the respondent and their local team; operated by their local campus; a state or regional resource; national resources; international government-funded; and commercial.

Average across responses: Percentage Use of CI type

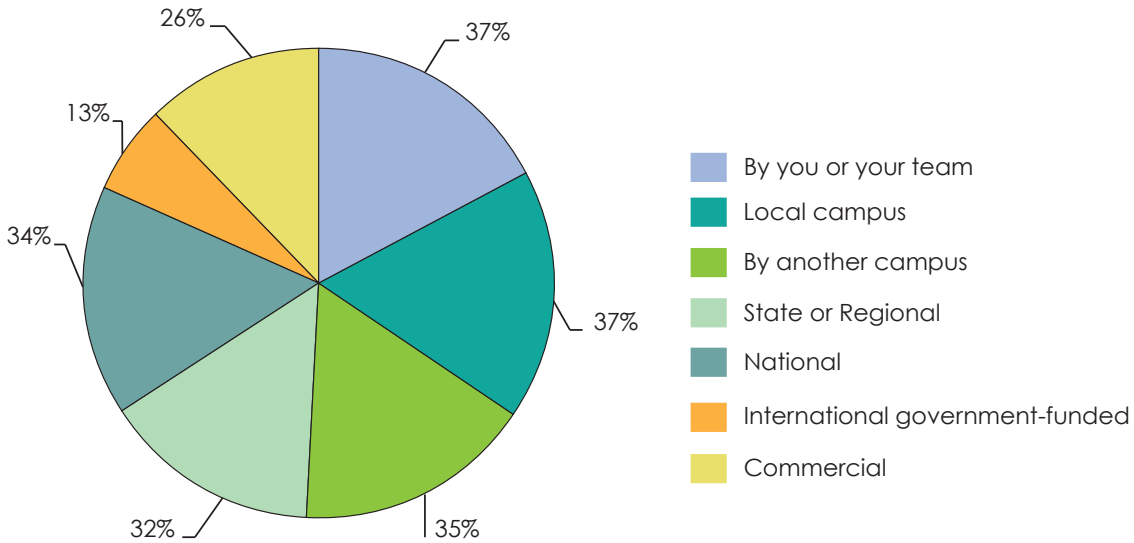


Figure A.1.2.

The respondents were then asked to characterize their usage of CI by an additional dimension of type of CI (data storage, simulation HPC/HTC, etc) and by the type of CI provider as in the previous question.

Question	You and team	Local campus	Another campus	State/ Regional	Natl.	Intl.	Commercial
Data storage	30.4%	20.7%	6.3%	1.9%	4.5%	1.3%	10.7%
Data movement	22.5%	15.0%	4.4%	1.8%	3.7%	1.5%	7.6%
Remote instrument access	10.6%	6.4%	2.4%	0.6%	1.0%	0.6%	1.7%
Workflows	5.1%	2.3%	0.6%	0.1%	0.5%	0.4%	1.7%
Collaboration	18.8%	14.1%	6.6%	1.1%	3.1%	2.2%	11.7%
Simulation HPC/HTC	13.3%	11.0%	3.7%	1.2%	4.9%	0.9%	1.3%
Large scale data analyses	13.0%	8.0%	2.9%	0.8%	4.1%	1.2%	2.2%
Visualization	17.9%	6.2%	2.3%	0.8%	3.0%	1.1%	4.3%

Table A.1.1.

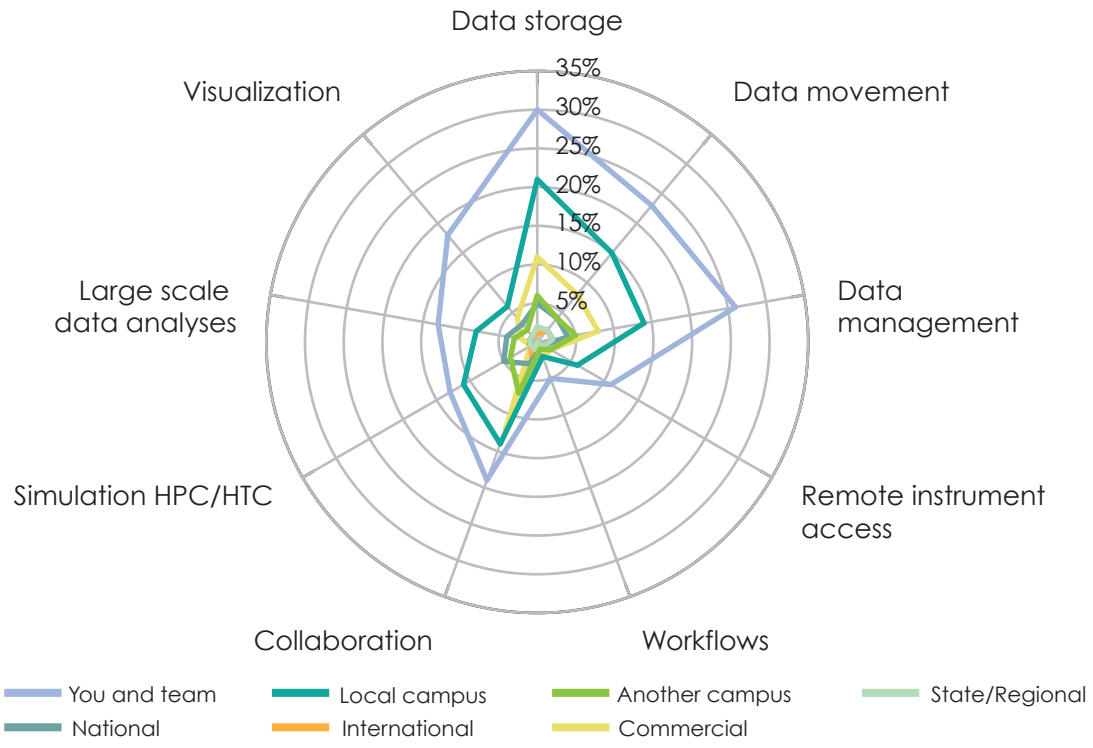

Data storage
Data movement
Data management
Remote instrument access
Workflows
Collaboration
Simulation HPC/HTC
Large scale data analyses
Visualization
35%
30%
25%
20%
15%
10%
5%
You and team
Local campus
Another campus
State/Regional
National
International
Commercial

Figure A.1.3.

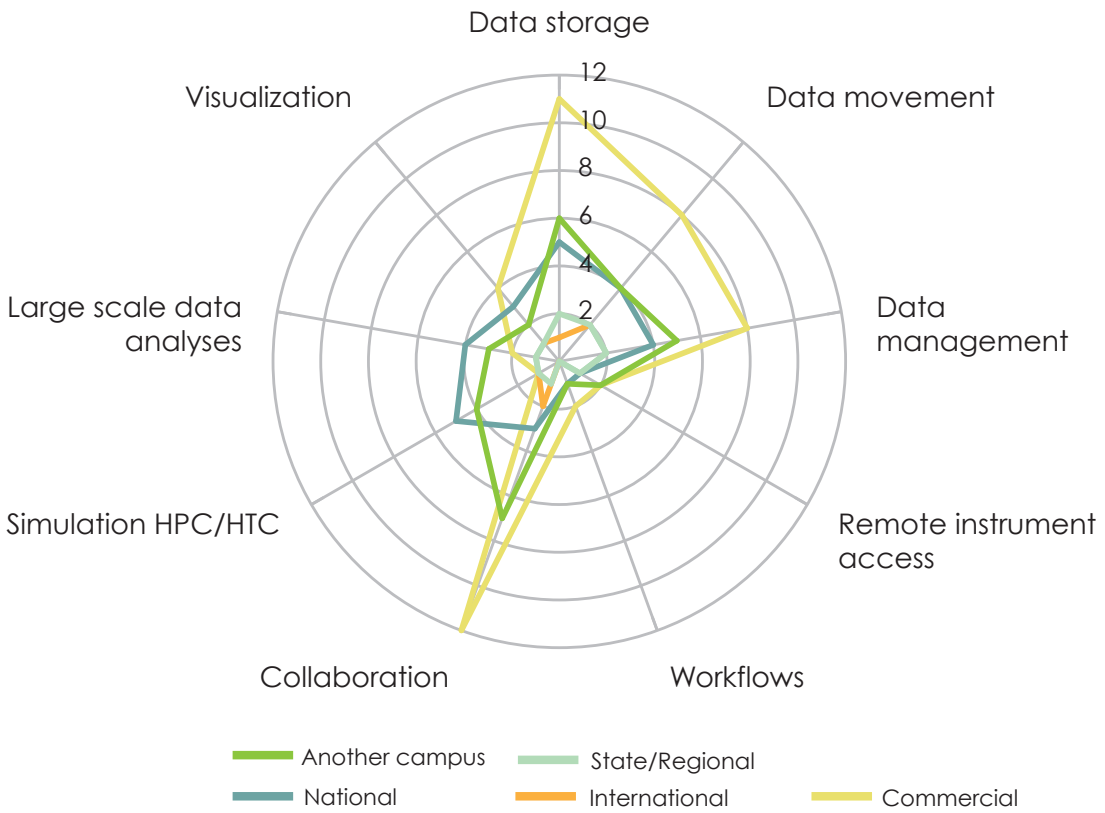


Figure A.1.4. This figure graphs the same data as A.1.3, however without the two largest data series (You and team; Local campus) so that usage characteristics of CI outside of the local campus can be more easily understood.

The respondents were asked to identify the methods by which they access CI. Answers included: Web browser and portal; SSH or other COTS command-line tools; graphical client applications; specialized middleware (Globus, Condor, Genesis II, etc); and other. We note that SSH is a common mechanism to access local, campus and national resources, however commercial services are highly concentrated with Web browser, portal, and graphical applications.

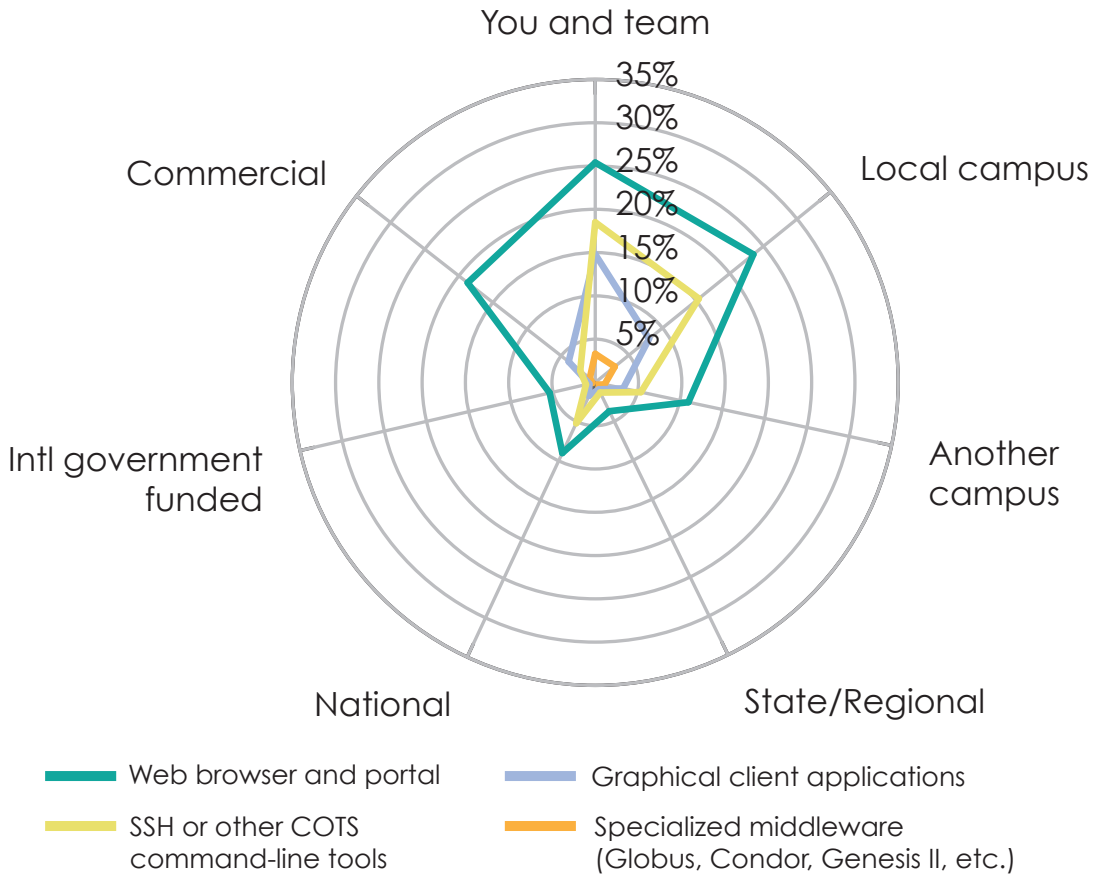


Figure A.1.5.

The respondents were asked about how they prefer to acquire support and help for their CI related efforts. Answers included: expert one-on-one consulting with operator staff or other expert; knowledge bases and other online documentation; user forums or other peer-to-peer communication; user advisory boards; workshops and training (on-line or in person); other.

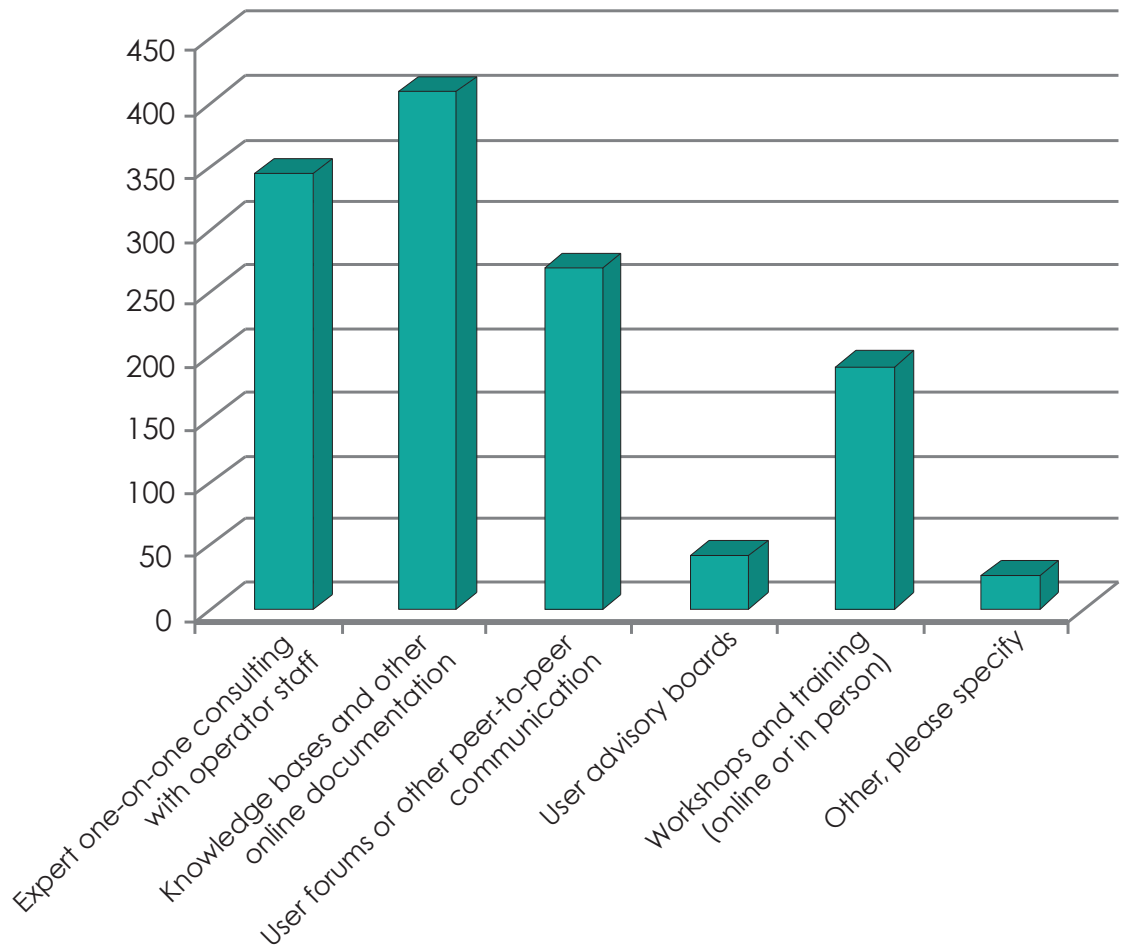


Figure A.1.6.

In the case of "Other," the respondents entered the following:

- Books
- Case studies
- colleagues' help
- Conferences like TGxy and SCxy
- expertise exchange visits
- feedback from very good experts of the particular systems, which is VERY HARD to get
- grad students
- in-house expertise
- knowledgeable people who respond to email
- library resources
- My husband.
- my RAs figure it out and tell me

- other users
- our own indisciplinary group
- peer networks (eg im and social networking)
- Productive graduate students
- public web sites
- quick responses to questions
- The HELP buttons
- Trial and error
- use cases, examples, tutorials
- well-defined user interface
- word of mouth

The respondents were asked about barriers to using CI. Specifically, they were presented with a list of seven potential barriers and asked to rank them from most important (1) to least important (7). The table below indicates that 51 respondents noted "allocation policies of remote resource" as the number 1 barrier to using CI, and that for 75 respondents, this particular barrier was not in their list of top 3 (i.e., 24+29+17+5).

Answer	1	2	3	4	5	6	7
Allocation policies of remote resource	51	30	27	24	29	17	5
Access control/ identity management issues (e.g., keeping track of accounts and passwords, different accounts and passwords	36	54	23	25	24	20	1
Ability to get support for remote resources locally	34	24	65	26	24	8	2
Ability to get support for remote resources from the resource provider	7	31	29	74	20	21	1
Local network throughput	17	16	16	11	56	55	12
Software compatibility with remote resource	18	23	19	22	30	61	10
Other (explain)	20	5	4	1	0	1	152

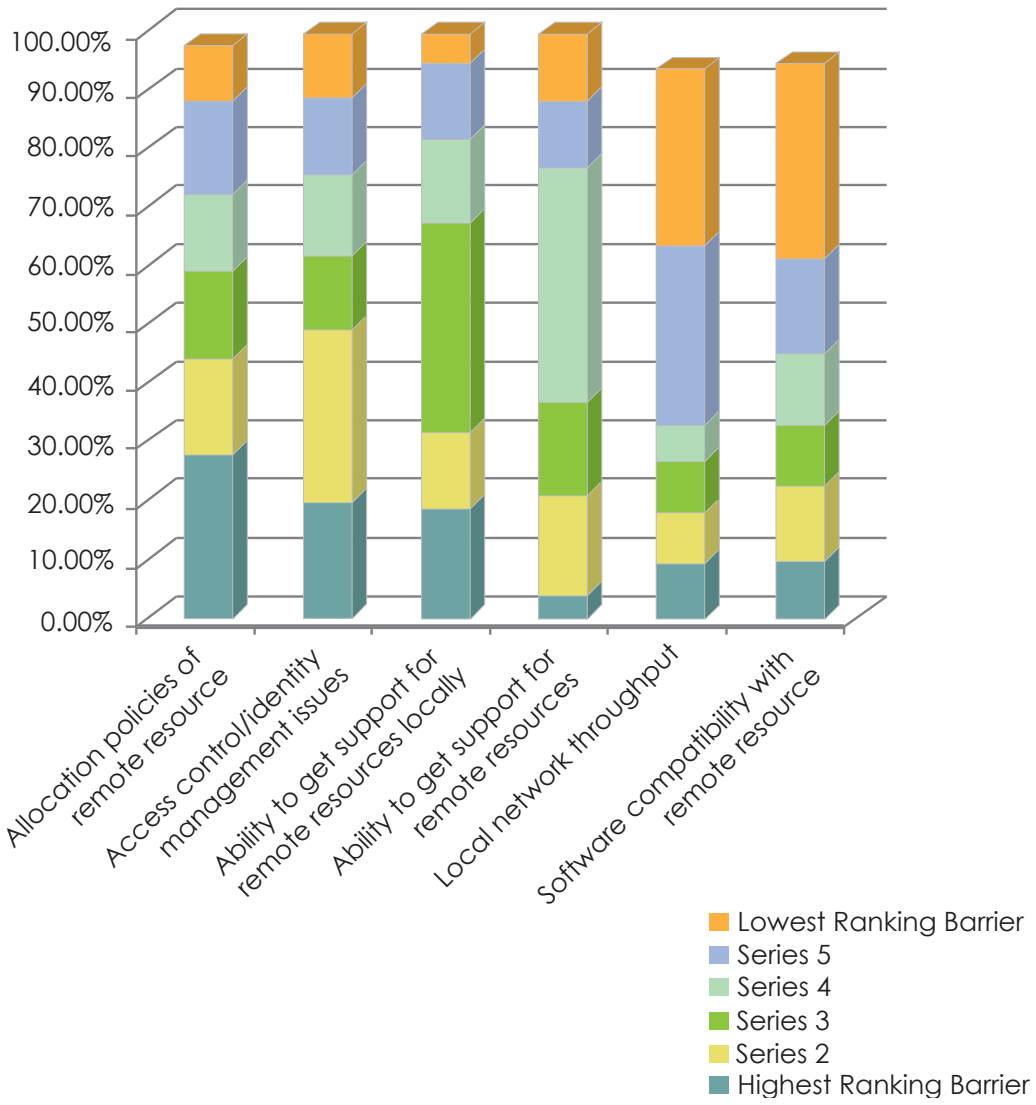

100.00%
90.00%
80.00%
70.00%
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
Allocation policies of remote resource
Access control/identity management issues
Ability to get support for remote resources locally
Ability to get support for remote resources
Local network throughput
Software compatibility with remote resource
Lowest Ranking Barrier
Series 5
Series 4
Series 3
Series 2
Highest Ranking Barrier

Figure A.1.7.

Table A.1.2.

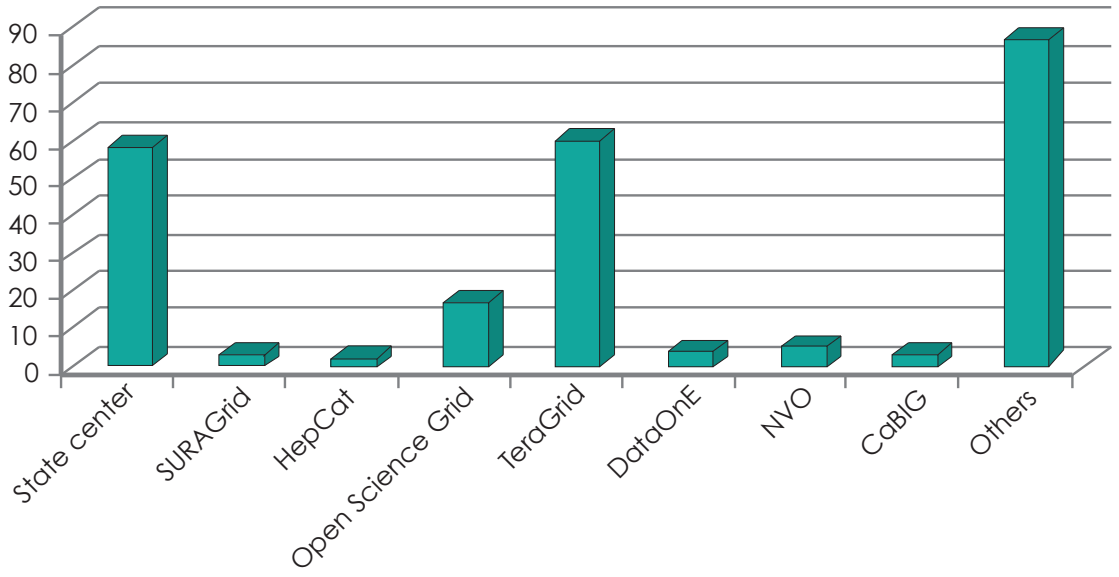


Figure A.1.8.

Respondents were asked about specific CI programs and if they are used.

Alabama Supercomputing Center
California
Cascades Volcano Observatory
CCNI
Center for Internet Augmented Rsrch and Asmnt
ChesapeakeBayNet
ESOGIS (NY)
GECO (Golden Energy Computing Organization)
Geographic Data Clearinghouse
Georgia GIS Clearinghouse
IACAT. NCSA
Indiana
Kansas - DASC,
Kansas Research and Education Network (KanREN)
Louisiana
Maryland
Mayaguez, Perto Rico
MCSR, Olemiss, MS
MHEC
Montana NRIS
NCAR/UCAR
NH
NHDOT
NM Encanto
NMGGI
NSCA
NY
Ohio Supercomputer Center
OR
PA
Pittsburgh Supercomputer Center
RENCI
San Diego
STC TRUST
TACC
Texas
TMSTEC
TXDOT
U of Florida
U of Utah
UMass Lowell's
USGS
Utah Education Network
Utah State U. IT
Vermont Advanced Computing Center
ZMATH, MATHSCINET

In the case of "Other," the respondents entered the following:

Finally, the respondents were asked if there was CI that they have tried but are not using. Only 10% of respondents selected "Yes," there is CI that they have tried but are not using. For those 10% of respondents, the fill in answers that they entered include:

- a cluster of computers (Beowulf)in my institution
- A high-performance computing cluster operated at the college level
- access grid and similar conferencing technology
- All of the grid stuff
- AMAZON
- Bluewaters
- caBIG
- Campus HPC cluster, National Super Computers (UCSD, Pitt)
- CIPRES portal

- clickers, podcast
- Clusters of the university research computing center
- Collaborative websites for projects-- Data Archives
- Community clusters
- Computing on national supercomputing centers such as the San Diego Supercomputer Center
- condor
- CVS based at Los Alamos National Labs
- Databases I have received training on.
- desktop sharing
- Document repository at partner institution. It's just easier to email necessary documents back and forth
- Due the problem of the firewall, our system cannot be shared freely with the community. We ended up using password request for accessing our information.
- Emulab testbed
- Euro Science Grid
- EVO
- Explored the option of TeraGrid but couldn't find a way to use it.
- for collaboration: DropBox, Google Wave, and a lot of other things I cannot remember
- Geon Neon archaeoinformatics
- Globus
- Google Apps
- grid-enabled or grid-accessible computational resources for bioinformatics
- HPC on campus; unfamiliar to me at this moment which particular cluster hardware/software is currently being used, though I have sought training and attempted to use it.
- HPCC cluster at USC
- http://www.phylo.org/
- I am semiliterate in this area. There are a lot of national databases but I am not trained in using them so I don't
- I have tried a highspeed IBM cluster at Brown University, I lack the systems support to install software on a parallel system
- IBM BlueGene
- Interactive video class for distance learning
- Internet2
- Kepler and Morpho (NCEAS products)
- "leadership class" computing centers
- Large File Sharing sites, Social Network Sites
- Local campus High Performance Computing Center.
- local Research Computing Center
- Local university resources

- Multiple resources on campus and nationally. Biggest issue is bandwidth and speed. This should be a priority issue. Important not only for data transfer but also for video communications.
- NAVDAT
- NCEAS database storage
- NCSA facilities at Urbana
- NCSU HPC
- Ocean Data View
- off-site data storage
- Ohio OSC
- Open Science Grid
- Other available processors at my Institution
- Planet Lab
- Remote instrumentation
- Remote supercomputing facilities (various locations)
- Resources at Sandia National Labs associated with LAMMPS
- Secure data storage at IU.
- Several servers placed at different locations (countries)
- still testing amazon cloud services, some collaborative tools - too complex and unreliable,
- Supercomputing Center
- SURA grid
- Teleconferencing/Videoconferencing
- TeraGrid
- University's high performance computing cluster
- Various collaboration options - wikis, google wave, skype
- Various grid computing systems
- Various kinds of user groups on Google. They keep changing them around.
- We had a Condor system set up on campus a few years ago. We haven't used it in a couple of years.
- We have tried the NASA high performance computing, but are not using it for any real computation because there is basically zero hard disk space available on the machines, and transferring data on and off is difficult.
- Webex

Appendix 1.2. Lessons Learned

Due to the timeframe of the NSF ACCI Task Force on Campus Bridging, there was insufficient time to implement a small-scale survey, analyze the results and then re-calibrate for a more comprehensive survey. As a result, we have learned a number of valuable lessons about the survey process itself and how we could have improved on the survey for the benefit of understanding the CI usage models of US researchers.

- Collect additional information about the respondents, specifically the domain of science and or NSF directorate. NSF sponsored activities cover a very broad range research, being able to correlate and include science domain as a dimension for the analyses would be very useful.
- Need to explore why researchers are not using CI more? A full 53% of the respondents indicated that they are not using CI "beyond their personal workstation or other CI operated by themselves or their team." Correlation with science domain and other questions could probe whether they simply are not aware that it exists, do not understand what is possible, are already operating their own significant CI, or is it truly not applicable to their research and scholarly efforts.
- Based on survey feedback, we believe special attention should be paid to not using terms or jargon that is unique to or most common in IT or NSF-OCI vernacular. Terms that cannot be avoided (e.g., CI) should be clearly defined as part of the survey.
- Datasets are a class of resource we overlooked (e.g., genome databases).
- True to our intention, we should emphasize that the survey is intended to gleam real knowledge about CI usage and requirements, and not just trying to justify CI.

Appendix 2. Submitted workshop papers

Workshop participants and members of the public were invited to submit white papers for consideration by workshop participants. Four white papers were submitted, which are included here in their entirety.

Appendix 2.1. Bridging Resources Using Lustre as a Wide Area Filesystem

Stephen Simms, Indiana University

Difficulties with data movement between resources and data management across administrative domains can create barriers for the process of scientific discovery. Indiana Universtity believes that the use of a wide area filesystem can be a very effective way to facilitate data sharing across distance and enable workflows spanning geographically distributed resources. The goal of this paper is to provide a few examples of how Indiana University has used the Lustre filesystem to span a single campus and bridge multiple campuses.

In 2005 the NSF granted $1.72 million to Indiana University for the construction of a storage facility, which we called the Data Capacitor. The goal of the system was to provide users with a high speed, high capacity filesystem for the short to mid-term storage of large research data sets. We chose the parallel distributed Lustre filesystem because of its scalability, speed, and ability to serve many clients. The Data Capacitor sits at the center of IU's cyberinfrastructure where it participates in all phases of the data lifecycle. As part of the grant we constructed 10Gb pipes to campus laboratories in order to permit a fast path between instruments and the Data Capacitor filesystem. We created a bridge between those campus resources and local cyberinfrastructure. At the same

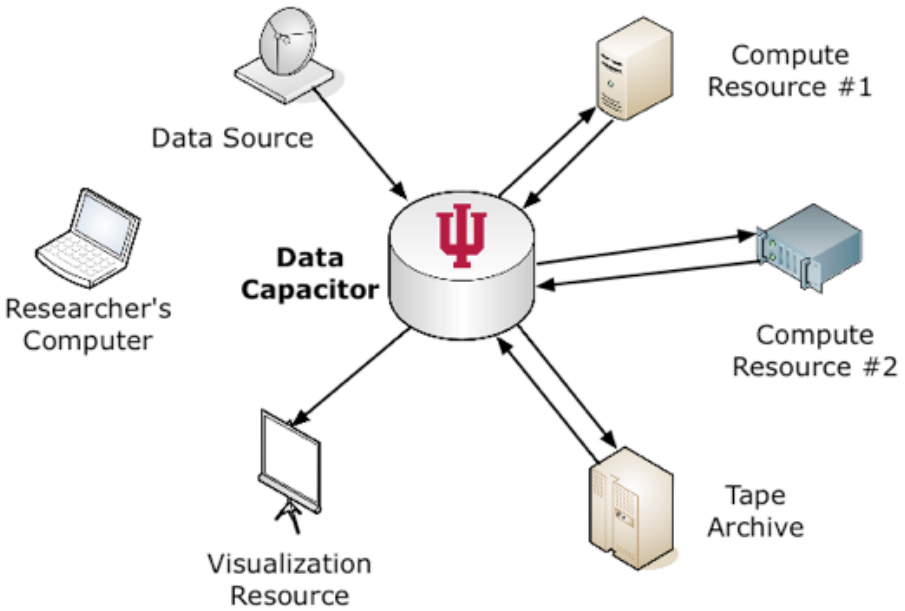


Figure 1. Data Capacitor spanning campus resources

time we created a bridge between campus resources and the national cyberinfrastructure that IU was providing as a TeraGrid resource provider.

Through a set of tests with Pittsburgh Supercomputing Center and Oak Ridge National Laboratory we discovered that using Lustre across the wide area network was not only possible, but could deliver excellent performance. From this data, we decided to try and expand the model that we had used locally. We wanted to include resources that were geographically distributed. At SC07 we entered the Bandwidth challenge to test and demonstrate the viability of using the WAN filesystem for computation across distance. We performed five different scientific workflows across distance (one spanning 5580 miles) and were able to saturate a 10Gb link, winning the competition.

For small specific projects with a limited number of participants it was possible to maintain a unified UID space. Users of the filesystem had the same UIDs across machines in order to insure that the ownership and permissions of files were consistent across domains. To extend beyond demonstrations and small projects, it was necessary to develop a scheme whereby UIDs could be mapped between heterogeneous domains.

IU solved the problem by developing a lightweight UID mapping scheme which allowed clients to use a standard Lustre distribution. In effect, users and remote administrators didn't have to worry about the mapping problem because it was solved on the server side maintained at IU. This development has enabled the filesystem to safely span multiple domains.

Bridging University of Kansas and IU

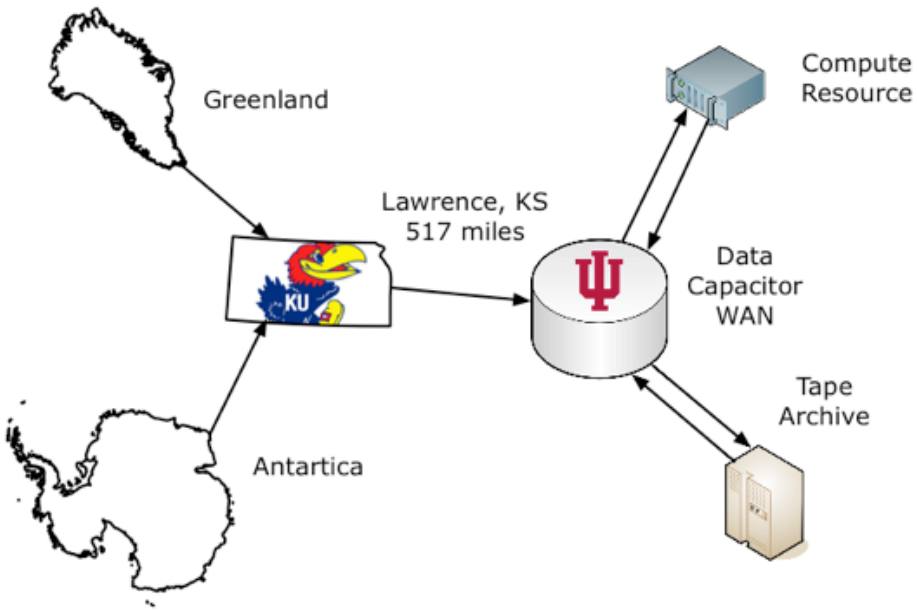


Figure 2. CReSIS Workflow

Researchers from the Center for Remote Sensing of Ice Sheets (CReSIS) at the University of Kansas contacted Indiana University in the Summer 2008 requesting supercomputing cycles to analyze their data. CReSIS researchers needed a way to move approximately 20 terabytes of data collected on the Greenland polar ice caps from Kansas to Indiana University. Once at IU, the data would be processed on IU's Quarry supercomputer. Through the use of a TeraGrid data allocation, CReSIS was able to use IU's High Performance Storage System (HPSS) to archive their results as well as their raw data.

The CReSIS research team needed a safe and reliable method to transfer and store the data. The Greenland data set was so large that CReSIS originally thought the best solution was to copy data onto USB drives and then physically ship them to Indiana for processing using a commercial carrier. Mounting the Data Capacitor in Kansas across Internet2, we were able to move all 20 terabytes to Indiana faster than they could have been copied to USB drives.

Here we have an example of a non-TeraGrid institution using the wide area filesystem as a bridge to TeraGrid resources. Researchers were able to collect, compute against, and archive data in a central location. The need to supervise complex data transfer mechanisms was eliminated.

Bridging the Texas Advanced Computing Center and IU

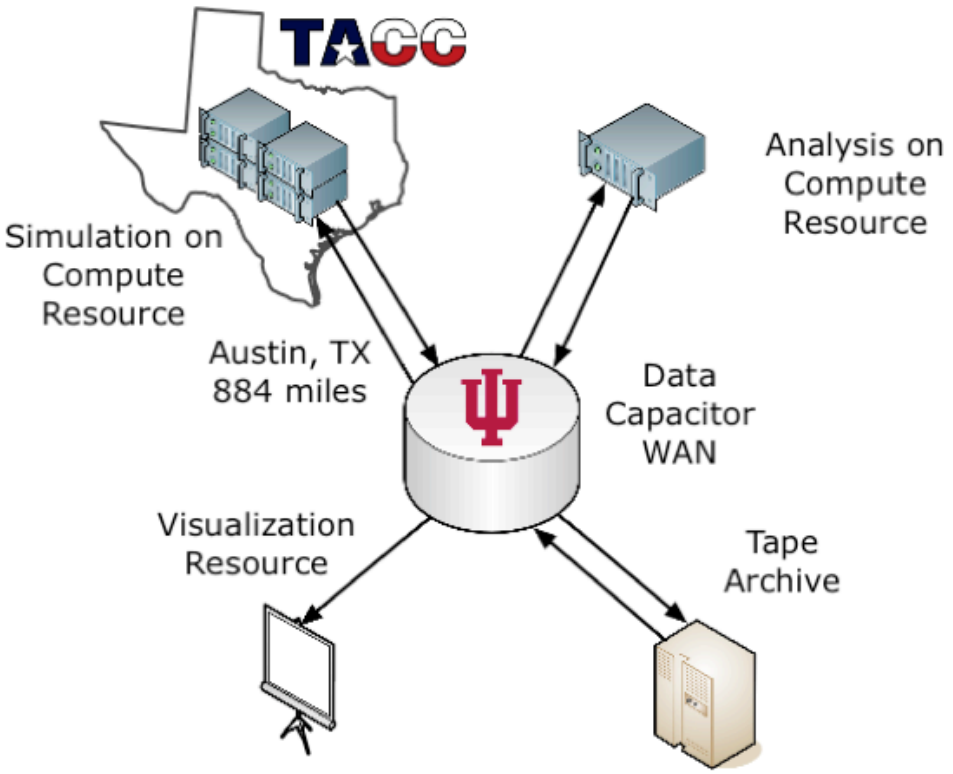


Figure 3. Dr. Chuck Horowitz's Workflow

Dr. Chuck Horowitz is a professor of physics at IU. His work involves simulating the behavior of matter in the crusts and interiors of neutron stars. This phase of matter is extremely dense and the laws of physics may behave differently at these extreme densities. Horowitz uses a molecular dynamics code to simulate several atomic species in the nuclear plasma. Early on, Dr. Horowitz was using the Data Capacitor to capture the output from the MDGRAPE-2 compute resources at IU for local visualization. With a TeraGrid allocation and the Data Capacitor's wide area filesystem, Dr. Horowitz was able to continue production at IU while performing additional larger scale simulations with the Lonestar cluster at the Texas Advanced Computer Center.

Here we have an example of the wide area filesystem creating a bridge between TeraGrid sites allowing Horowitz the ability to easily aggregate and compare simulations from multiple sources without having to use cumbersome data transfer mechanisms. Additionally, no performance differential was measured between runs against Lonestar's local filesystem and the Data Capacitor wide area filesystem.

Bridging PSC, NCSA, and Mississippi State University

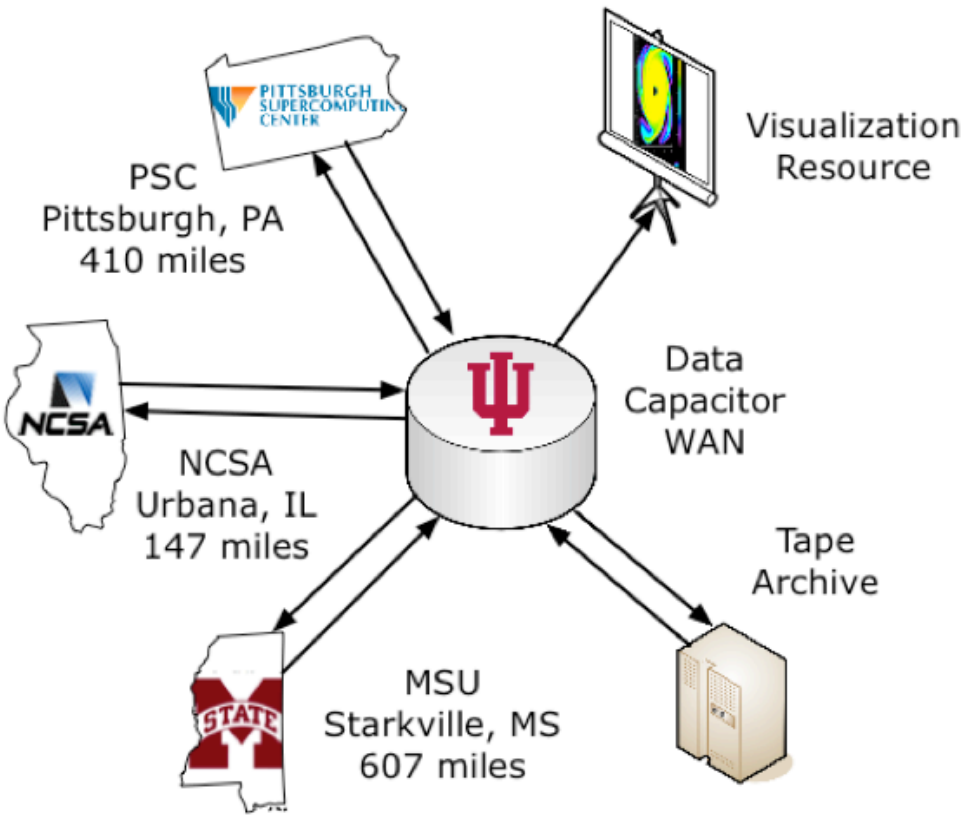


Figure 4. Dr. Richard Durisen's workflow

Dr. Richard Durisen is currently using the Data Capacitor and the TeraGrid to simulate gravitational instabilities in protoplanetary discs to study the origins of gas giant planets. Durisen's research team uses Pople, a shared-memory system at PSC as well as NCSA's Cobalt to produce more than 50 TBs of simulation data. The second part of their workflow is to perform an analysis of the simulation data produced. This step was performed at Mississippi State University on their Raptor cluster.

By using the Data Capacitor to bridge IU, PSC, and NCSA, Durisen's team can see results as if they were happening locally. Should there be a problem with the run or the input parameters it is possible to stop the simulation and correct the problem without wasting valuable CPU time. Additionally, being able to see the simulation unfold can give the researcher insight into the next run (or set of runs) to be performed.

By using the Data Capacitor to bridge IU and MSU, Durisen's team was able to take advantage of some free compute cycles that were made available. Normally, this would require the researcher to transfer the data between sites, but the Data Capacitor made this step unnecessary.

Conclusion

Using the Data Capacitor as a central filesystem has permitted data sharing across distance and simplified workflows that would have previously been more complex and involved significant data transfer. When data is centrally located it can make data management significantly easier.

The Data Capacitor also provides a resource for storing large data sets that might overwhelm a smaller institution. For example, the scratch filesystem at MSU was 5TB in size. Durisen's team would have had to fill that filesystem 10 times over to complete their analysis.

One might expect significant performance degradation using a wide area filesystem, however this was not always the case. In each case where data was produced across distance there was no performance differential between local and remote mounts. In the case where data was analyzed at MSU, there was a difference between the local and remote filesystems of 40%. However, this does not take into account the considerable amount of time that would be required to transfer 50 TB of data to the local filesystem.

Appendix 2.2. Position Paper on Campus Bridging Software and Software Service Issues

Rion Dooley, Texas Advanced Computing Center

Integrating disparate resources both vertically and horizontally is an extremely difficult task. There are significant technological challenges when bridging different system architectures as well as political problems when bridging across authoritative domains. In this position paper, we briefly touch on some of the most desirable characteristics of a cyberinfrastructure (CI) that bridges campuses. This paper is in no way meant to be exhaustive, nor is it meant to provide deep insight into the topics it lists, but rather it is meant to serve as an guide for further discussion and investigation based on our experiences using and integrating campus, national, and international resources. The remainder of this paper is grouped into two sections. The first section is non-technical and gives 10 rules of thumb for developing any bridging CI. The second section is more technical and covers some architectural and design guidelines for the CI software as a whole.

Section 1: Rules of Thumb

1. Start focused on software and gradually shift that focus to hardware over the duration of a purchasing cycle. Local universities cannot compete in economies of scale, but they can partner to leverage them. With virtualization, partnering today means something completely different than partnering 10 years ago. It's worth the paperwork to make it happen.
2. Software should virtualize the underlying resources into a black box for the majority of users. Just as the batch scheduler is the interface familiar to users today, so too should the CI provide a common, intelligent, familiar interface for users tomorrow.
3. Learning curves are healthy. We will learn more if we stop avoiding the struggle. As long as the CI and message are consistent, users will learn, adopt, and adapt.
4. CI can and should be a playground for strategic partnerships. It should be flexible and sufficiently mature for RP of all sizes to establish transient partnerships with each other.
5. The goal of CI is synergy. Anything less cannot be called a success.
6. Good CI leads to innovation. Innovation leads to change. Change should feed back into the CI. If no real innovation occurs from anyone using the CI, the CI isn't really usable.
7. We are the United States of America. That's a good model for understanding institutional dynamics. Institutions are made of people. People need to feel important. Any CI that does not allow people to retain a sense of autonomy will fail.
8. Build the playground. Define the rules. Give people a chance to follow them, and reward the ones that do.

9. Make sure technology trickles down to the local level. A flexible CI can go big or small. Make it easy to use for local systems that are not ready to play right now, and when they are, they will play nice.
10. Pain is an effective motivator. Seminars and tutorials are not sufficient to bring people up to speed. Engaging users on their turf, in their environment allows them to have their own struggle-victory story...and share it with you. Those relationships are priceless in building bridges of any kind.

Section 2: The Software

System Level

A bridging CI should provide a common abstraction layer for users to do their work. This starts at the systems level, working with vendors make sure that the basic building blocks of accounting, monitoring, job management, networking, security, and storage are in place before building up a CI. This is a gap in modern middleware. Lots of effort is spent accounting for everything happening through the middleware, but little attention is paid to the things happening outside the middleware. This leads to inconsistencies for users moving between the middleware and the underlying system.

Building upon the core services mentioned above is an event-driven middleware stack that's developed and maintained as part of the national CI. We specifically mention that the middlware should be event driven so that it allows for both push and pull interaction from users. This is important because different users have different use cases and it's is important to enable both models of access through the middleware.

Security

The underlying security model should be built upon existing standards such as Kerberos and LDAP, and exposed through widely accepted cross-domain solutions such as OAuth that allow for trust relationships, single sign-on, and the ability to generate bi-directional audit trails on demand. Building out custom authentication frameworks is simply not necessary when widely supported solutions exist. Even if commercial licenses are required for such tools, it will be less expensive than funding a 5 year effort to custom design something. Adopting an existing, supported solution will also reduce the integration burden on the CI developers and RP.

Breadth

The CI should also be sufficiently flexible to decouple the concept of an organization from the underlying resources. As systems and data requirements continue to grow at divergent paces, it will become increasingly important to see organizations appear who specialize in specific tasks such as data storage or instrumentation. If a CI cannot represent such single-purpose entities, it will be difficult to create meaningful partnerships using the CI.

Maintainability

This software stack is community driven, but controlled by a persistent development team that ensures the software exhibits a high quality of service and benefits from relevant technology insertion over time. The role of this layer of the CI is overall system management. It is an abstraction layer to interact with the core system services in a uniform way. This layer should be available as a suit of command line tools as well as programmatic and web service APIs. We make no comment on the physical implementation other than it should be sufficiently fast and responsive to be utilized in a web interface without causing confusion to the end user. Also, it should work in such a way that users can move from the CI to the local resource without losing track of their jobs or their work. This requires intelligent identity management and accounting, but that needs to be conducted by the underlying system and monitored by the CI.

Applications

The role of the CI is not to support end user applications, but to enable them. This means that it is not the role of the CI or support team to build, benchmark, and maintain applications. That responsibility falls on the shoulders of the RP, vendors, and owners of the individual codes. Some of these applications will be made public for general use. In these situations the CI should support some notion of software lookup and discovery service.

Marketing

Finally, it is important to remember that several attempts to build a useful CI have been made before. History has shown that without flexibility, engagement, and education, they fade away due to lack of adoption. Marketing is critical. It begins at the national level and continues at every level down to the local campuses. Everyone involved needs to have a clear understanding of what the CI is, what it does, how it will help them, and what their role in the larger collaborative will then be.

Such an education and outreach effort can significantly lower the barrier to entry for coming generations of users, however it will not get the traction it needs to achieve national synergy without support from university administrators and vendors. Bridges are only useful when people cross them. If we encouraged people to take a different path to their destination, there's no point in building the bridge. Thus, it is the responsibility of influential leaders to point people to path of least diversion.

Appendix 2.3. Barriers to Inter-Campus Cyberinfrastructure

John-Paul Robinson, University of Alabama at Birmingham

Two initiatives addressing the challenges faced in bridging campuses with the national CI are the Open Science Grid (OSG) and SURAgrid. Both organizations have had difficulties fostering campus adoption of technologies that address resource sharing at a large scale. There have been successes, however, under specific circumstances. Understanding the elements to these successes helps highlight barriers to campus bridging. To overcome these barriers, we need to address organizational problems that inhibit adopting available resources.

Almost ten years ago, the NMI Testbed exposed participants to resource sharing technologies emerging from large-scale national collaborations. The technologies promised all manner of new paradigms. Back then it was called sharing compute cycles across the grid. Today, we'd call this cloud computing and shared data centers.

When the testbed ended, many of the participants wanted to continue these efforts and banded together under the umbrella of campus grids. You could say, we intentionally wandered into a land of promise, hopeful, that leveraging the interfaces of the national computing infrastructure on campus would provide our computational scientists a smooth transition to more power as their demand grew. Things changed along the way, though: everyone has become a computational scientist and the value of a utility computing extends far beyond HPC.

What started as a loosely coupled outgrowth of the NMI Testbed eventually grew into SURAgrid, a campus-to-campus grid community that provides a valuable engagement point for participants to learn about grid interconnection and HPC operations from peers who share the goal of expanding access to computational resources. This volunteer effort eventually established a membership and governance structure as the need to coordinate and focus efforts increased. The community developed a four year strategic plan to define shared goals and provide direction on a timeline through 2012. SURAgrid is a campus-to-campus, CI-focused virtual organization sustained by the voluntary efforts of its members.

An open community is key to building expertise and interest in new paradigms. In addition to building this CI-focused community, one of the most measurable successes of SURAgrid has been an HPC cluster contract in partnership with IBM. The contract provided aggregated buying power to what otherwise would have been smaller, isolated campus acquisitions. The contracts also reserved 20% of the acquired compute cycles to be used as a shared compute pool for SURAgrid. A key factor to the success of this program is that acquiring a cluster is a very tangible driver for a campus and its role as a production system motivates support for local users accessing the cluster.

These local interfaces, however, have been very different from the "grid" interfacing set up for external access and this difference has made it difficult to share compute cycles. With the core mission addressed, the federated interfaces become secondary and have proven difficult to maintain, especially when hurdles remain and only limited time or expertise is available to address

these shortcomings. Services critical to local operations are readily maintained and disparate interfacing discourages resource sharing.

OSG has built a scalable infrastructure by providing a comprehensive platform for its adopters. The platform addresses the need for a consistent user experience and a straight-forward implementation. As with cluster acquisition, the key behind a successful adoption is a strong driver that makes the OSG resources a critical component of the research enterprise. OSG user communities have a strong, application-oriented focus. This application focus serves to hide the differences between sites from users. In many implementations OSG users also do not have to learn two models of resource interaction; a Condor scheduling abstraction provides consistent access to local and remote resources. Services critical to local operations are readily maintained and application-oriented interfacing encourages resource sharing.

OSG has faced the same difficulties as SURAgrid when engaging the campus. OSG has been easiest to incorporate at sites that can adopt the OSG platform completely or that have the expertise needed to integrate resources. It's relatively easy for users to adopt a platform when their existing practices don't conflict with the proposed model. Most users will adapt to what ever services are supported locally. If those services can incorporate distributed resources smoothly they will be happy to use them. If remote resources require separate processes, adoption will be very low.

The problems common across participation in our campus CI build outs, SURAgrid efforts and OSG explorations have largely been a question of labor. For example, the ability to support applications on a resource is often limited by system administrator time. Many sites implicitly assume a local system administrator will install shared applications. The ability to separate roles into application maintainers and system maintainers helps. Unfortunately, this is often a local solution and varies from site to site. Good processes in one community aren't necessarily available or known to another. This type of overhead between systems makes most people avoid using distributed resources. Changing the focus from sharing computers to sharing applications would significantly ease resource sharing. This requires trust in a new role, the application administrator, and demands curation and record keeping to validate configurations when needed. Maintaining infrastructure to support this work is currently left to individual sites and adds to the burden of supporting users. At other times, knowledge of a resource integration solution has existed at one site only to be absent from the next. The state and utility of campus systems is extremely variable and severely limits inter-campus resource sharing. Certainly tools do exist to address these problems, however, they often grow out of domain specific initiatives or require re-engineering work flows. Campus service providers are all too often lightly staffed and the ability to help any one group in great detail is limited. If users can't adopt new frameworks themselves they will simply stick to what's working. The single biggest challenge in campus bridging is the limited ability to share labor effort across organizational boundaries.

The applications focus of the cloud paradigm is a promising driver for large-scale resource sharing. Identifying business needs for the campus to adopt and support the infrastructure will be crucial to its success. While the core interconnecting technologies to bridge the campus exist, much

work remains to build curated collections of research applications. Most campuses do not have the resources to support these efforts alone. Providing a framework to reduce construction and maintenance costs of this platform is the key to harnessing manpower embedded at the campus. The organizing models demonstrated by the open source community provide guideposts for harnessing distributed labor and building trust across administrative boundaries; Linux kernel development has shown shared resources can be built to address the competing business interest of contributors; and new platforms like Google's Android continue to reveal the business value of building a coherent system held in common by its contributors. Building bridges requires coordinated effort and a commitment by the campus to incorporate the shared resources as primary components of their operations. Without or ability to harness incremental labor contributions across organizational boundaries, campus bridging will forever remain on the fringes of infrastructure.

Appendix 2.4. Campus Bridging as an Exercise in Cultural Bridging

Dan Fraser, Production Coordinator, Open Science Grid

One strategy for Campus Bridging is to first enable a Grid within the campus and then extend the Campus grid by connecting to (or integrating with) external grids such as TeraGrid, EGEE, or the Open Science Grid (OSG). While it may be tempting to assume that one could simply implement one of the existing grid infrastructures on a campus, in practice this approach is problematic – many existing grid infrastructures require a significant effort both to operate and even more importantly they require a steep learning curve to use. One way of exploring this problem is to note hat each of the aforementioned grids has created its own culture that includes: terminology, access patterns, usage patterns, security, networking, and operational styles, together with a sophisticated set of processes for authentication, authorization, monitoring, accounting, and data management. Unfortunately, every campus already has its own culture with each of these items and it is rare when a campus has more than a few of these cultural items that are the same as in the grid world. Hence, Campus Bridging is not a technological problem nearly as much as it is an exercise in "cultural bridging".

Bridging a cultural gap means going back to basics and first trying to understand the campus culture. Many questions come to mind: How many resources are available on the campus? How sophisticated are the system administrators? What would be the advantages of combining these resources into a campus grid? Where are the campus users that are (or could be) interested in large scale computing? What kinds of problems do they have? Are they interested in massively parallel computing or high throughput computing (lots of single processor jobs)? How big of a role does the campus IT play? Can one build on the infrastructure that is already in place?

Suppose one considers a single potential campus grid user. How does that user scale from whatever resource they are currently using to take advantage of a larger infrastructure? What might that infrastructure look like so as to make the user transition as easy as possible?

Some of the most valuable information that has been learned in the grid computing world is not which technology to adopt, but where the problems are. For example, the most consistent complaints we hear from the grid user community, especially new users, invariably involve security, and grid certificate management. Therefore it is worth asking whether grid certificates are required for a campus grid in a campus environment that already has a healthy security model? Can campus users use their own campus security model for a campus grid?

For additional examples, it is helpful to consider a particular grid infrastructure. With over 80 active sites and a usage pattern that is 100% grid based (e.g. users do not login and submit interactive jobs), the Open Science Grid (OSG) is one successful model to learn from. The OSG for example is as much about sociology as it is about technology and process. To consider a positive example, one might ask: What are the strategies that have most helped users in overcoming site differences across

the grid? On the OSG the answer is overlay technology such as Glide-in WMS[1]. Hence it makes sense to ask whether or not glide-in overlay technologies could be useful on a Campus Grid.

One way that the OSG has enabled multiple sites to function together is by providing a common layer of middleware that is divided into compute elements and storage elements[2]. Do such components make sense in a campus environment? The OSG also recommends certain common naming conventions such as $OSG_WN_TMP and $OSG_APP that provide access storage that users can depend on at each site. What types of naming conventions would be useful in a Campus Grid?

Bridging cultures requires asking many questions. In the above section we have attempted to identify a few of these questions with the hope that in researching answers to these questions we can begin to create effective solutions for campus grids that will be welcomed by the campus communities who adopt them and introduce far fewer problems than the number they solve.

1 https://twiki.grid.iu.edu/bin/view/Documentation/JobSubmissionComparison
2 http://www.opensciencegrid.org/About/Learn_About_Us/OSG_Architecture

Appendix 3. Prior work

A number of prior works with regards to user cyberinfrastructure requirements, use cases and evaluation fed into our workshop and the CI user survey:

- TeraGrid held a workshop for its user community in 2006 coordinated by Zimmerman and Finholt [7]. This workshop is similar in its goals with our workshop in that it sought to identify from a scientist's perspective what aspects of the TeraGrid was working well at that time.
- The final report of the NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences [8] contains a categorization of software for data-oriented cyberinfrastructure that serves as input for our user survey.
- The "Open Grid Services Architecture Use Cases" from the Open Grid Forum [9] lays out a number of cyberinfrastructure use cases. While many of these are from the point of view of a cyberinfrastructure deployer, some are user-centric and helped shape our user survey.
- "Perspectives on Distributed Computing" from Lisa Childers, et al [10] reports on interviews with 30 cyberinfrastructure users, cyberinfrastructure deployers and cyberinfrastructure developers. While somewhat focused on Globus-specific use cases, it served as input to our user survey.
- The 2007 "Study of User Priorities for e-Infrastructure for e-Research (SUPER)" by Newhouse et al [11] was a very similar effort in the U.K. to our workshop in that their goals were to identify current and near-term gaps in the U.K. e-Science infrastructure through both an online survey and user interviews. We leveraged their survey and interview questions in our user survey.
- TeraGrid is producing a Usage Modalities document by D. Katz, et al, which was presented by Daniel S. Katz as a work in progress at our workshop and served as input to discussions.
- In 2008, the EDUCAUSE Center for Applied Research conducted a short study of cyberinfrastructure resources and practices among the EDUCAUSE membership [12]. This survey covered broad technologies areas (e.g., data storage and management, HPC resources) and asked respondents to rank current and future importance of these areas across a number of academic areas (science and engineering, other disciplines, creative activities, and teaching and learning). This work was used to shape our CI user survey.
- In 2008, UCSD conducted a campus user survey, which was reported on in "Blueprint for a Digital University: A Report of the UCSD Research Cyberinfrastructure Design Team (4/24/09)" [13]. We unfortunately learned of this work only after we designed our survey.

Appendix 4. Workshop agenda

The workshop was held in Denver, CO August 26-27th at the Sheraton Downtown Denver Hotel, a location chosen due to the availability of flights from both coasts and as part of a strategy to distribute the Campus Bridging workshops across the country (the first workshop was in Indianapolis and the third will be in California). The workshop was one and a half days, starting on the 26th and ending at noon on the 27th. The detailed agenda of the workshop follows.

August 26th

- 9:00 am: Welcome, introductions and review of workshop goals (Von Welch)
- 9:30 am: Presentation of survey results and discussion (John McGee)
- 10:30 am: Break
- 11:00 am: Daniel S. Katz: "TeraGrid Usage Modalities" (Introduction: Bill Barnett)
- 11:30 am: Discussion of bridging methodologies. Can we agree on small number (~6) of broad bridging use cases to help organize subsequent discussion? (Lead: John McGee)
- Noon: Lunch
- 1:00 pm: Continue discussion of bridging methodologies from before lunch.
- 1:45 pm: Effective strategies for CI User Support. "Given fix resources for user support, how would you focus those resources between the various approaches (e.g., knowledge bases, help desk, expert consulting, on-line documentation, training)?" Each presenter has 5 minutes to give their views (no slides) and then discussion.
    - Amit Majumdar
    - Dan Frazier
    - Kim Dillman
    - Moderator: John McGee
- 3:00 pm: Break
- 3:30 pm: "What are the biggest points of pain for researchers in using the CI on-ramps today?" Each presenter has 5 minutes to give their views (no slides) and then discussion.
    - Andrew Grimshaw
    - Miron Livny
    - Ken Klingenstein
    - Moderator: Von Welch
- 5:00 pm: Adjourn
- 6:00 pm: Dinner

August 27th

- 8:00 am: Campus Perspective: "What role should campuses be playing in the international CI ecosystem? What should the CI ecosystem being doing for campuses?" Each presenter has 5 minutes to give their views (no slides) and then discussion.
    - Jim Pepin
    - David Walker
    - Gary Crane
    - Moderator: Bill Barnett
- • 9:30 am: Break
- 10:00 am Discuss "parking lot" (unresolved) issues. (Lead: John McGee)
- 11:00 am: Summarize and discuss workshop conclusions. (Lead: Von Welch)
- Noon: Wrap-up and conclude

Appendix 5. Student participants

To broaden impact and include an educational component, four scholarships were awarded for student participation in the workshop. The scholarships included travel and lodging expenses along with a small stipend to cover additional expenses. Applicants were required to submit a CV highlighting their past and planned experience with CI. Six applications were received and the organizing committee selected the following four students:

- Rajendar Kanukanti, LONI (Graduate student in Computer Science)
- Jonathan Morrison, Indiana University (Graduate student in Computer Science)
- Vani Panguluri, Southern University (Graduate student in Computer Science)
- Man Luo, University of North Carolina (Graduate student in Pharmaceutical Sciences)

Appendix 6. References

1. Indiana University Pervasive Technology Institute. Campus Bridging. Available from: http://pti.iu.edu/campusbridging [cited 21 Mar 2011]
2. NSF 10-015. Dear Colleague Letter: Cyberinfrastructure Framework for 21st Century Science and Engineering (CF21). http://nsf.gov/pubs/2010/nsf10015/nsf10015.jsp
3. Cyberinfrastructure Software Sustainability and Reusability Workshop Final Report. C.A. Stewart, G.T. Almes, D.S. McCaulay and B.C. Wheeler, eds., 2010. Available from: http://hdl.handle.net/2022/6701 [cited 21 Mar 2011]
4. EscinetWiki. Campus Grids. Available from: http://wikis.nesc.ac.uk/escinet/Campus_Grids [cited 21 Mar 2011]
5. National Cancer Institute. Knowledge Centers: Domain Experts. Available from: https://cabig.nci.nih.gov/esn/knowledge_centers [cited 21 Mar 2011]
6. TeraGrid. Campus Champions. Available from: https://www.teragrid.org/web/eot/campus_champions [cited 21 Mar 2011]
7. Zimmerman, A. and Finholt, T. TeraGrid User Workshop Final Report. Collaboratory for Research on Electronic Work, School of Information, University of Michigan, 2006.
8. Berman, F. and Brady, H. Final Report: NSF SBE-CISE Workshop on Cyberinfrastructure and the Social Sciences. 2005.
9. GFD-I.029. Open Grid Services Architecture Use Cases. Open Grid Forum, 2004.
10. Childers, L. Liming, L., and Foster, I. "Perspectives on Distributed Computing: Thirty People, Four User Types, and the Distributed Computing User Experience," Argonne National Laboratory Technical Report ANL/MCS/CI-31, September 2008.
11. Newhouse, S., Schopf, J., Richards, A. and Atkinson, M. Study of User Priorities for e-Infrastructure for e-Research (SUPER). UK e-Science Technical Report Series, November, 2007.
12. Sheehan, M. Higher Education IT and Cyberinfrastructure: Integrating Technologies for Scholarship (Research Study, Volume 3). Boulder, CO: EDUCAUSE Center for Applied Research, 2008.
13. UCSD Research Cyberinfrastructure Design Team. Blueprint for the Digital University. April 24, 2009.

Appendix 7. Workshop presentations

Workshop on Campus Bridging Software and Services

Welcome and Goals

Von Welch

Acknowledgements

- NSF
- Note takers:
  - Ray Sheppard
  - Jonathan Morrison
- Logistics, Survey, Printed Materials, more…
  - Dale Lantrip

Welcome to Students

Workshop Background

- Six task forces established by NSF Advisory Committee for Cyberinfrastructure (ACCI).
- Goal is to develop and establish Cyberinfrastructure Framework for 21st Century Science and Engineering

- HPC
- Software and Tools
- Work Force Development
- **Campus Bridging**
- Data
- Grand Challenges

Workshop Goals

- First, some working definitions…

(Proposed) Definitions

- Cyberinfrastructure (CI)
  - From NSF Cyberinfrastructure Vision for 21st Century Discovery:

"Cyberinfrastructure integrates hardware for computing, data and networks, digitally-enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools."

(Proposed) Definitions

- Campus Bridging:

Use of local (under user's control) CI in coordination with remote CI. Where 'remote CI' can be CI elsewhere on campus, another campus, regional CI, national CI, international CI, commercial CI.

(Proposed) Definitions

- Services:
  - Defined broadly.
  - Not just software services.
  - Includes user support, consulting, etc

Workshop Goal

- Try to answer the following...
  - What sorts of bridging are users doing?
  - What software and services are working well to enable bridging?
  - What is missing? What can't users do they want to?
  - What could be better? What are the pain points?
  - What will users want to do in the near future (<5 years) they aren't/can't today?

Agenda: Thursday

- Two presentations:
  - Survey of CI usage
  - Cyberinfrastructure Usage Modalities on the TeraGrid
- Discussion on usage modalities
- Panel: Effective strategies on CI User support
- Panel: CI Pain points
- Dinner

Agenda: Friday

- Breakfast @ 7:15, start @ 8
- Panel: Campus Perspectives
- Parking Lot issues
- Wrap-up by noon
  - Box lunches available

Question, comments, suggestions?

Over to John McGee

Survey results

before we start . . .
From previous section: CI Definition

From the March-2009 report of the summer-2008 joint Educause CCI WG and CASC workshop, "Developing a Coherent Cyberinfrastructure from Local Campus to National Facilities: Challenges and Strategies"

Cyberinfrastructure consists of computational systems, data and information management, advanced instruments, visualization environments, and people, all linked together by software and advanced networks to improve scholarly productivity and enable knowledge breakthroughs and discoveries not otherwise possible.

NSF ACCI CBTF
Campus Bridging Technologies Workshop

Survey on Cyberinfrastructure-Enabled Bridging

http://campusbridging.iu-pti.org/agenda

Denver, CO; 8/25/2010

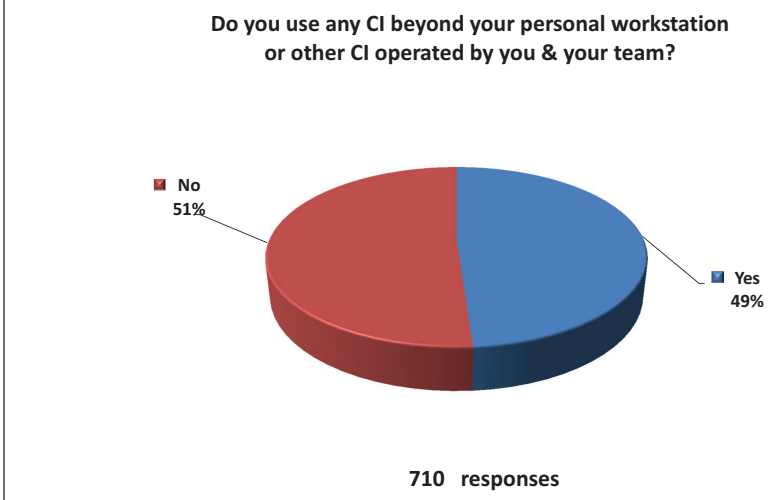
John McGee, Dale Lantrip, Von Welch, Craig Stewart

Goals for this session

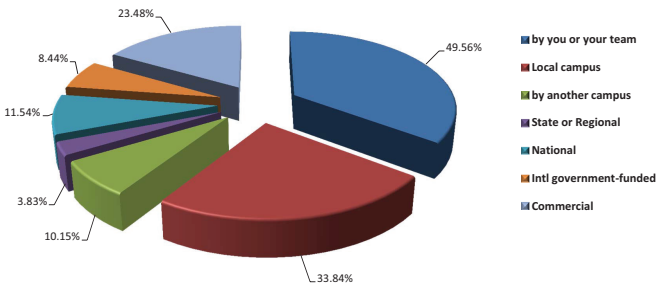
- Understand the data acquired thus far
- Scaffolding for discussion
- Thoughts on the survey process

Survey Details

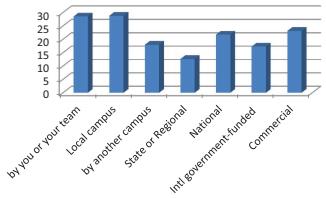
- More complete details in printed materials accompanying this workshop
- 5000 researchers selected randomly from a list of 34,623 NSF principal investigators (PIs)
- an ongoing survey initiated on 5 August 2010. A snapshot of results have been preliminarily gathered for the purpose of presentation at this workshop



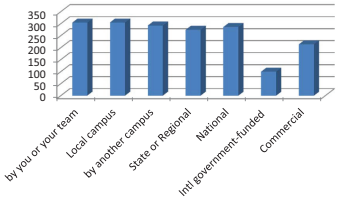
Percentage Use of CI type: average across responses



Std Dev



Responses

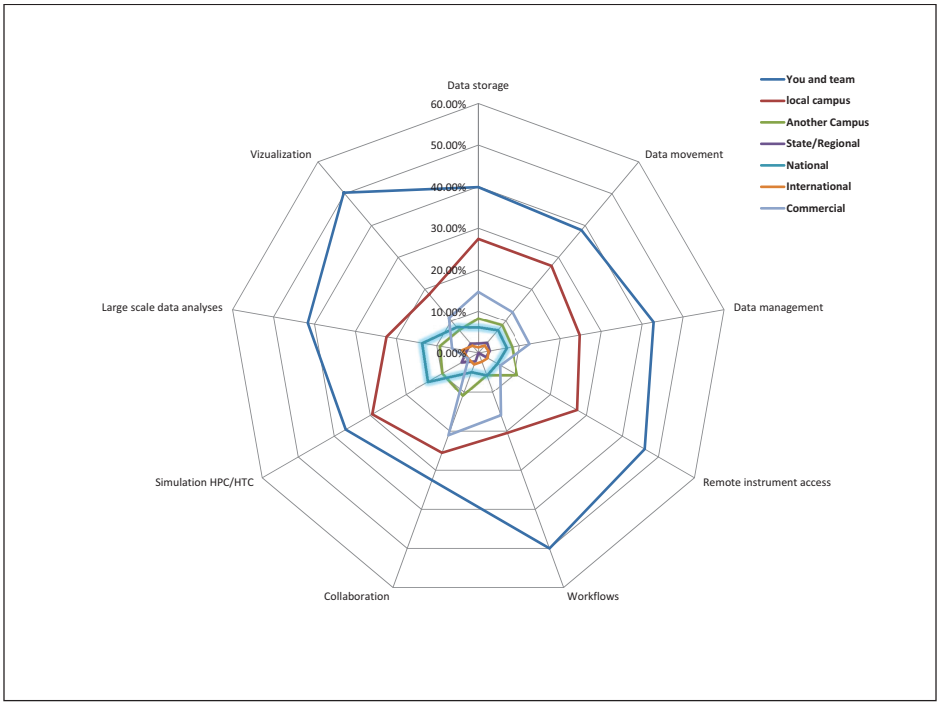


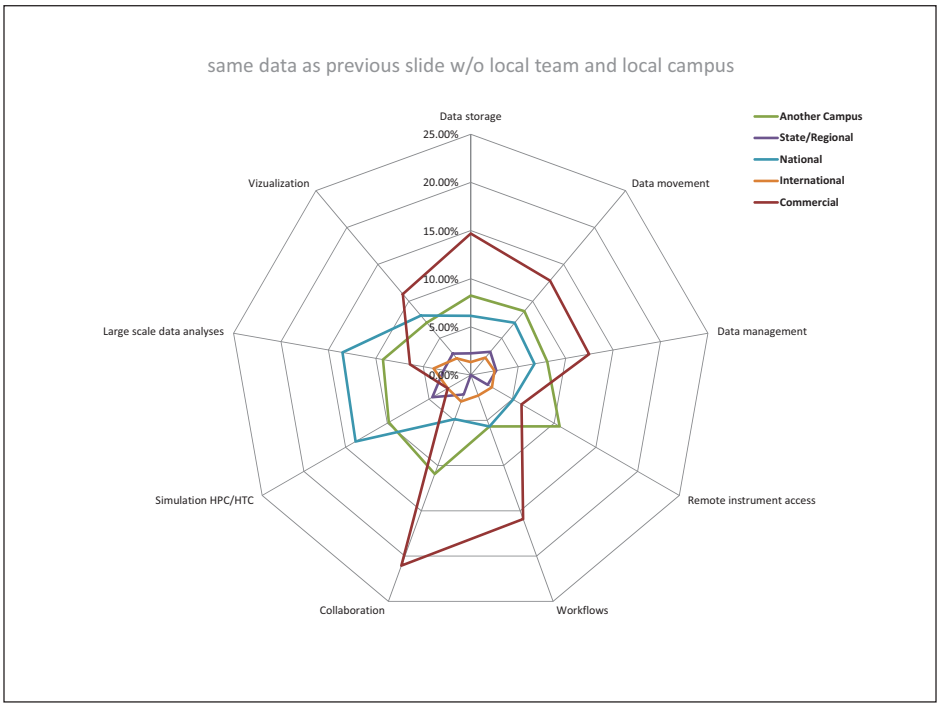
Survey on Cyberinfrastructure-Enabled Bridging

What type of functionality do you obtain from the different types of CI? Please check all boxes in the matrix that apply.

	Data storage	Data movement	Data management	Remote instrument access	Workflows (define)	Collaboration	Simulation (high performance computing or high throughput computing)	Large scale data analyses	Vizualization
Resources operated by you or your team	☐	☐	☐	☐	☐	☐	☐	☐	☐
Resources on your local campus	☐	☐	☐	☐	☐	☐	☐	☐	☐

#	Question	you or your team	local campus	another campus	state or regional	natoinal	internati onal	commerc ial	response s
1	Data storage	39.88%	27.44%	8.25%	2.25%	6.15%	1.35%	14.69%	667
2	Data movement	38.58%	27.36%	8.66%	3.15%	7.09%	2.36%	12.80%	508
3	Data management	42.80%	24.76%	8.06%	2.69%	6.72%	2.50%	12.48%	521
4	Remote instrument access	46.19%	27.41%	10.66%	2.03%	5.08%	2.54%	6.09%	197
5	Workflows	50.00%	20.45%	5.68%	0.00%	5.68%	2.27%	15.91%	88
6	Collaboration	32.55%	25.54%	10.92%	2.14%	4.87%	2.92%	21.05%	513
7	Simulation HPC/HTC	36.81%	29.45%	9.82%	4.60%	13.80%	2.76%	2.76%	326
8	Large scale data analyses	41.64%	22.42%	9.25%	2.85%	13.52%	3.91%	6.41%	281
9	Vizualization	50.32%	18.39%	7.10%	2.90%	8.06%	2.26%	10.97%	310


Data storage
Data movement
Data management
Remote instrument access
Workflows
Collaboration
Simulation HPC/HTC
Large scale data analyses
Vizualization
60.00%
50.00%
40.00%
30.00%
20.00%
10.00%
0.00%
You and team
local campus
Another Campus
State/Regional
National
International
Commercial

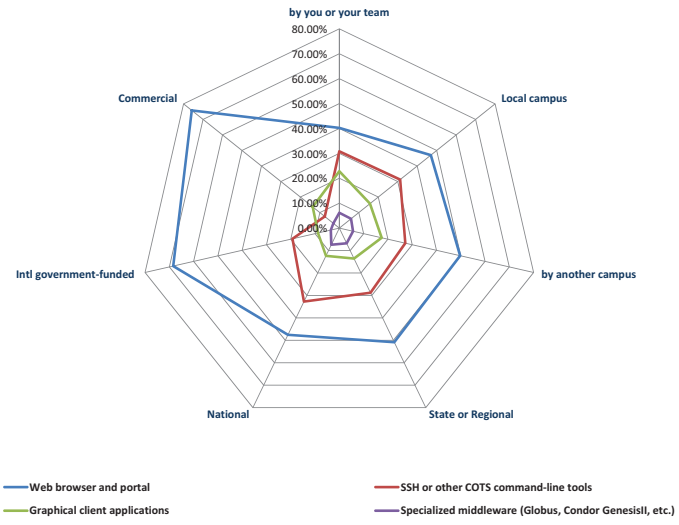

same data as previous slide w/o local team and local campus
Data storage
Data movement
Data management
Remote instrument access
Workflows
Collaboration
Simulation HPC/HTC
Large scale data analyses
Vizualization
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
Another Campus
State/Regional
National
International
Commercial

What OTHER types of functionality do you obtain from the different types of CI?

note: selected comments only

- Web hosting, IT infrastructure, security, backup
- Scientific publishing/editorial work
- Data discovery
- Software development and version management
- Grading resources
- Digital libraries, information retrieval
- Basic access to resource to assist others in debugging and using our software

What methods do you use to access CI?

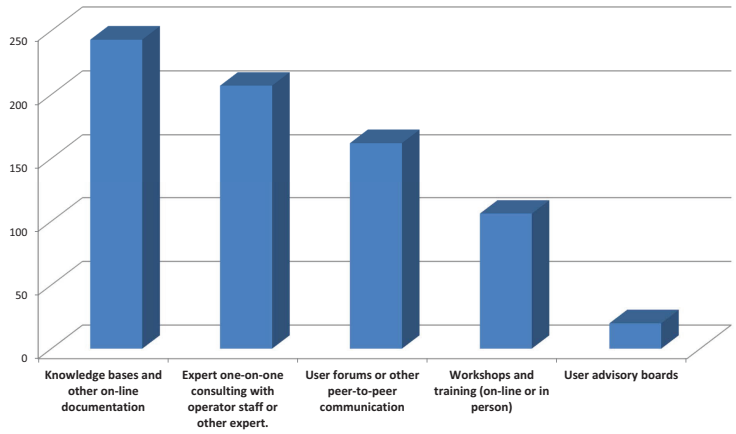


What OTHER methods do you use to access CI?

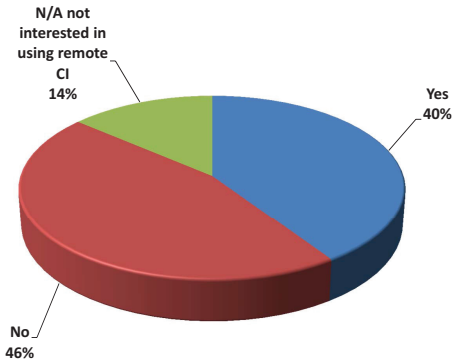
note: selected comments only

- FTP
- SQL
- SVN
- VNC
- Custom software
- Cloud, EC2, Azure
- BOINC

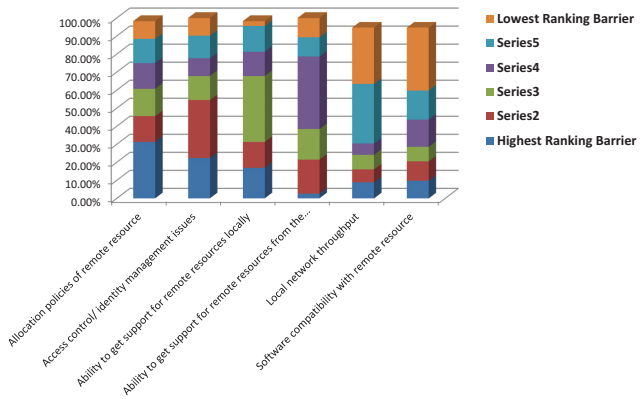
Which do you find useful to support your use of remote CI?



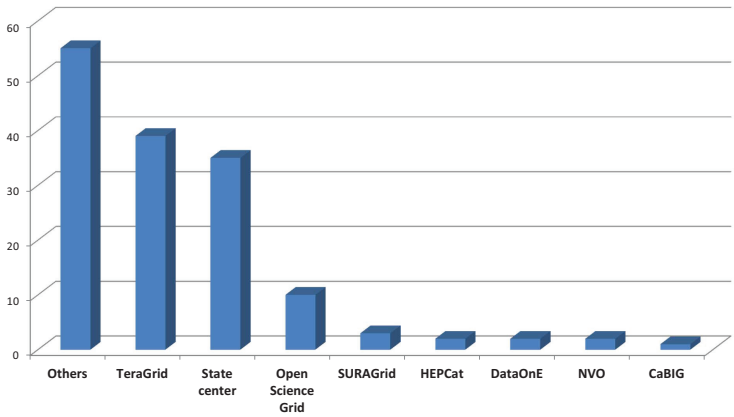
Are there barriers to your using remote CI?



Answer (ranking of barrier)	1	2	3	4	5	6
Allocation policies of remote resource	31.25%	14.29%	15.18%	14.29%	13.39%	9.82%
Access control/ identity management issues	22.32%	32.14%	13.39%	9.82%	12.50%	9.82%
Ability to get support for remote resources locally	16.96%	14.29%	36.61%	13.39%	14.29%	2.68%
Ability to get support for remote resources from the resource provider	2.68%	18.75%	16.96%	40.18%	10.71%	10.71%
Local network throughput	8.93%	7.14%	8.04%	6.25%	33.04%	31.25%
Software compatibility with remote resource	9.82%	10.71%	8.04%	15.18%	16.07%	34.82%



What state, regional, national, or international CI facilities are you using now?



What state, regional, national, or international CI facilities are you using now? (Other)

Amazon Web Services
ArcGIS online
Argonne Natl Labs
ARIn databases
Blackboard
CAMERA at Calit
CCT @Uky
CIPRES
DETER testbed
DoD centers
Encyclopedia of Life
ESG
Flybase
Gbif
GenBank
Google Earth
Internet 2
iPlant
Knowledge Network for Biocomplexity
LLNL
Los Alamos National Lab
Magellan at ANL
MIDAS cluster
MSPnet
NASA Ames
NCAR
NCBI
NCEAS
NEOS
NERSC
NIST
NLR
OSC
national data sources hosted by universities
PCMDI
Planetlab
Polar Information Commons
Starlight
USGS National Map
various research software packages

TeraGrid Usage Modalities

Slides prepared and presented by
Daniel S. Katz
d.katz@ieee.org
Director of Science, TeraGrid GIG
Senior Fellow, Computation Institute, University of Chicago & Argonne National Laboratory
Affiliate Faculty, Center for Computation & Technology, LSU
Adjunct Associate Professor, ECE, LSU

Ideas and work are collaborative with TeraGrid, specifically including: David Hart (SDSC -> NCAR), Chris Jordan (TACC), Amit Majumder (SDSC), J.P. Navarro (UC/ANL), Warren Smith (TACC), Von Welch (NCSA -> Von Welch Consulting)

What is the TeraGrid

- World's largest (arguably) distributed cyberinfrastructure for open scientific research, supported by US NSF
- Integrated high performance computers (>2 PF HPC & >27000 HTC CPUs), data resources (>3 PB disk, >60 PB tape, data collections), visualization, experimental facilities (VMs, GPUs, FPGAs), network at 11 Resource Provider sites
- Freely allocated to US researchers and their collaborators
  - Researchers request time, peers review and determine merit, TG staff fit requests to resources
- Mission:
  - DEEP: provide powerful computational resources to enable research that can't otherwise be accomplished
  - WIDE: grow the community of computational science and make the resources easily accessible
  - OPEN: connect with new resources and institutions
- Integration: Single: portal, sign-on, help desk, allocations process, advanced user support, EOT, campus champions

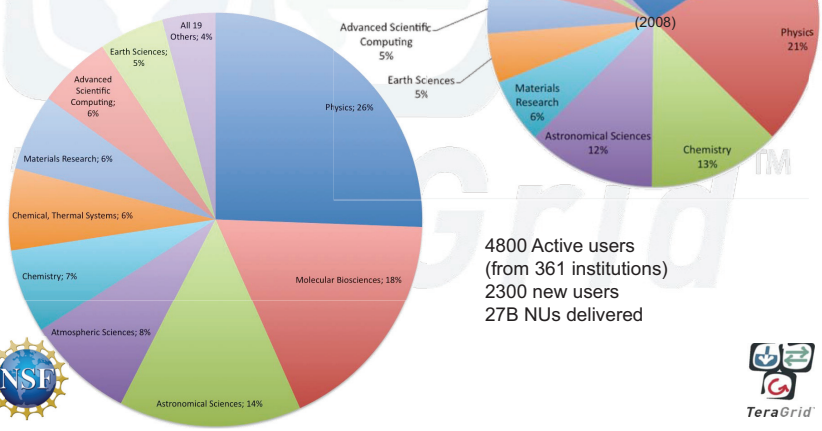
Governance

- 11+ Resource Providers (RPs) funded under separate agreements with NSF
  - Indiana, LONI, NCAR, NCSA, NICS, ORNL, PSC, Purdue, SDSC, TACC, UC/ANL (& GaTech)
  - Different start and end dates, different goals, different agreements, different funding models (sometimes within a RP)
- 1 Coordinating Body – Grid Infrastructure Group (GIG)
  - University of Chicago/Argonne National Laboratory
  - Subcontracts to all RPs and six other universities
  - 8-10 Area Directors
  - Working groups with members from many RPs
- TeraGrid Forum with Chair

TeraGrid -> XD Future

- Current RP agreements end in March± 2011
  - Except track 2 centers (current and future)
- Most of TeraGrid XD (eXtreme Digital) starts in April 2011
  - 2 services (TAS, TIS) started in April 2010
  - Era of potential interoperation with OSG and others
  - New types of science applications?
- Current TG GIG continues through July 2011
  - Allows four months of overlap in coordination
  - Some overlap between GIG and XD members
- Blue Waters (track 1) production in 2011

Who Uses TeraGrid (2009)



4800 Active users
(from 361 institutions)
2300 new users
27B NUs delivered

How One Uses TeraGrid

Slide modified from Dane Skow and Craig Stewart

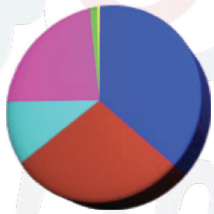
Science Gateways

- A natural extension of Internet & Web 2.0
- Idea resonates with Scientists
  - Researchers can imagine scientific capabilities provided through familiar interface
    - Mostly web portal or web or client-server program
- Designed by communities; provide interfaces understood by those communities
  - Also provide access to greater capabilities (back end)
  - Without user understand details of capabilities
  - Scientists know they can undertake more complex analyses and that's all they want to focus on
  - TeraGrid provides tools to help developer
- Seamless access doesn't come for free
  - Hinges on very capable developer

Slide courtesy of Nancy Wilkins-Diehr

How TeraGrid Was Used (2006)

Usage Modality	Community Size (rough est. - number of users)
Batch Computing on Individual Resources	850
Exploratory and Application Porting	650
Workflow, Ensemble, and Parameter Sweep	250
Science Gateway Access	500
Remote Interactive Steering and Visualization	35
Tightly-Coupled Distributed Computation	10



Usage Modalities

- Why do we want this data?
  - We want to know what our users are trying to do
    - Perhaps what they ideally want to do
    - At least how they are using our infrastructure now
  - We can use this information to develop new tools, improve existing tools, change operational practices and policies
- How was this data obtained in 2006?
  - Piecemeal at best
    - Including some data gathering, some guessing, some word-of-mouth, etc.

Usage Modality Requirements

- Define a space of modalities, with each orthogonal modality as a dimension, with a set of possibly values
- Must be able to measure each value, ideally directly using the TeraGrid infrastructure, at least by inference from user intent
  - Use units that are common or translatable (core-hour, NU, TB/yr)
- Must be able to tie measurement of an activity (user, job, file, etc.) in one dimension to measurement of same activity in other dimension(s)
- Caveat
  - Fairly limited in use of CI currently – focused on running compute jobs, small amount on storing and moving data, small amount on human expertise, little-to-nothing about sensors, experimental facilities, etc.

Current Draft Set of Modalities

- User intent
- When-to-run
- Submission-mechanism
- Resources
- Job coupling
- Support
- Level of program development

Usage Intent

- Definition
  - Why the user is doing this
- Values
  - Production
  - Exploration/porting
  - Education
- How to measure
  - Ask the user at the time of their allocation request
    - Estimate fraction of each value you plan
  - Multiply actual runs by these fractions
- Issues
  - Not very accurate

When-to-run

- Definition
  - When the user's activity should (needs to) start
- Values
  - Batch
  - Interactive
  - High Priority
  - Reservation
- How to measure
  - Local job scheduler measures everything – all batch by default
  - Tools/local job scheduler measure everything else
- Issues
  - Potentially many tools that need to be modified to count activities

Submission-mechanism

- Definition
  - How the user's activity is started
- Values
  - Command line
  - Grid tools
  - Science gateways
  - Metascheduler
- How to measure
  - Tools report this to TGCDB
- Issues
  - Science gateways use grid tools underneath
  - Metascheduler is a grid tool?

Resources

- Definition
  - What resources are needed to "run" the activity
- Values
  - 1 HPC resource, multiple HPC resources, 1 HTC resource, visualization resource, data-intensive resource, archival storage resource, multi-site storage resource, non-TG resource
- How to measure
  - Pull from TGCDB directly
- Issues
  - Incomplete?
  - Not yet clear what information is useful here

Job coupling

- Definition
  - How is this job (activity) related to others?
- Values
  - Independent (e.g., single job)
  - Independent but related (e.g., an element of parameter sweep)
  - Tightly coupled (e.g., a part of distributed MPI or component application, jobs that must run simultaneously)
  - Dependent (e.g., an element of a workflow, jobs that depend on other jobs)
- How to measure
  - Ask users to estimate their fraction of the four types as part of their allocation request, multiply usage by these fractions
- Issues
  - Not very accurate

Support

- Definition
  - Are we (TG) providing special support to the user? (Advanced user support is requested by the user as part of their allocation request, then peer reviewed, and possibly granted.)
- Values
  - Advanced support is being provided
  - Advanced support is not being provided
- How to measure
  - AUS is tracked in TGCDB – just pull this data
- Issues
  - Utility of this?
  - Other support we should measure?

Level of program development

- Definition
  - How much program development the user has done
- Values
  - Custom (some work done by user)
  - Commodity (no work done by user)
- How to measure
  - Check path to binary executable?
  - Build table of exceptions by hand?
- Issues
  - Custom is not very specific
    - Probably, some additional values would be useful
  - Even though this is perhaps too simplistic, it's already hard to measure

Going further

- Do these modalities make sense in other venues?
  - Campuses? Other national/international infrastructures?
- Measuring them is much easier if supported by tools
  - Common needs help encourage tool developers to support us
- Does this lead to the need for new tools that need to be developed?
  - Or common policies across different infrastructures?

Final Thoughts

- Every HPC system is unique
  - Moving HPC applications without virtualization is extremely difficult
  - Virtualization hurts performance too much for HPC applications
- Is this true of other resources – that they are unique?
- Do we want to hide this or expose it?
- Maybe worthwhile to think about:
  - What parts of CI can be made common with policies/tools?
  - What parts can't?
    - What do we do about this?

Workshop on Campus Bridging Software and Services

Findings and Recommendations

Draft Findings

Findings

- CI is as much about people as technology
  - People required for sustained infrastructure.
  - Hardware is cheap; lack of qualified staff a big problem
- Training and education is key and campuses can contribute to training and education of both users and operations staff
  - User population less programming-literate and/or used to different methodologies than in past
  - E.g. Uva CS 101 for Grad students (taught by CC)
- Differences between access methodologies new generation of users are used to and what CI provides
  - E.g. GUI vs CLI, Unix vs MacOS/Windows
  - Social mediazation
  - Need to educate users in these modalities (e.g. UVA programming class)…
  - And need to shift CI usage modalities

Findings

- Use of CI requires trust from users.
  - Users leery to trust CI not under their control
  - It becomes critical to their science (papers, Nobel prize, etc.)
- Users very loath to change once they have something that works
  - Especially to something with different usage modality
- Users experience drop-off in non-functional attributes as they move from their local CI.
  - Reliability, Availability, Dependability, Predictability, Usability
  - Commercial CI has caught on due to it addressing these issues
  - E.g. The campus firewall boundary
- Wordsmith on last thought:
  - CI needs to balance between reliability versus risk, capability, flexability as appropriate for community
  - Manage risk to optimize cost-benefit trade-off (will be different for enterprise)

Findings

- There is a lack of bridging within campuses between administrative computing and research computing
  - Do we want to build this bridge or built a moat? They are optimized for different things.
    - Fixing this bridge may go a long way to addressing inter-campus bridging
- Campus administration is generally putting more emphasis on administrative computing. Outreach to CIOs, VPRs, and other campus leadership is needed.

Findings

- Campus bridging is increasingly enabling scholarship.
  - Dovetailing of campus and non-campus CI.
  - Is about easy/seemless transversal of boundaries of control.
  - As much about connecting to other researchers as it is with connecting with CI.
- Measuring impact of campus bridging is an issue
  - Measuring total effort on research computing is hard

Findings

- NSF leadership would be of great value, needs to be firm
  - Many other priorities on campuses
  - Coordination with other agencies would help
  - Leadership may be effective even without complete funding.
- Need more incentives for operationally-focused CI
  - Necessary to establish user confidence
  - Review process should include evaluation if existing CI is being appropriately leveraged
  - Need to focus evaluation on effectiveness rather than speed, novelty, research
  - Addressing how it contributes to infrastructure
  - Addressing sustainability

Findings

- Interoperability of support infrastructure can go a long way. E.g.:
  - Ability of support staff to find each other across domains, local expertise
    - E.g. Campus champions and other CI expertise – needs to be general and not technology/project specific
  - Ticket systems to allow coordination between projects and campus staff
  - Job submission systems
  - Build, run-time environment
- Bringing expertise across domains a challenge
  - Don't know/trust users in other domains
  - Don't know who to ask

Findings

- Resources are difficult to discover
  - Resource policies are difficult to discover
  - Lack of visibility
- Smaller institutes, science communities difficult to discover

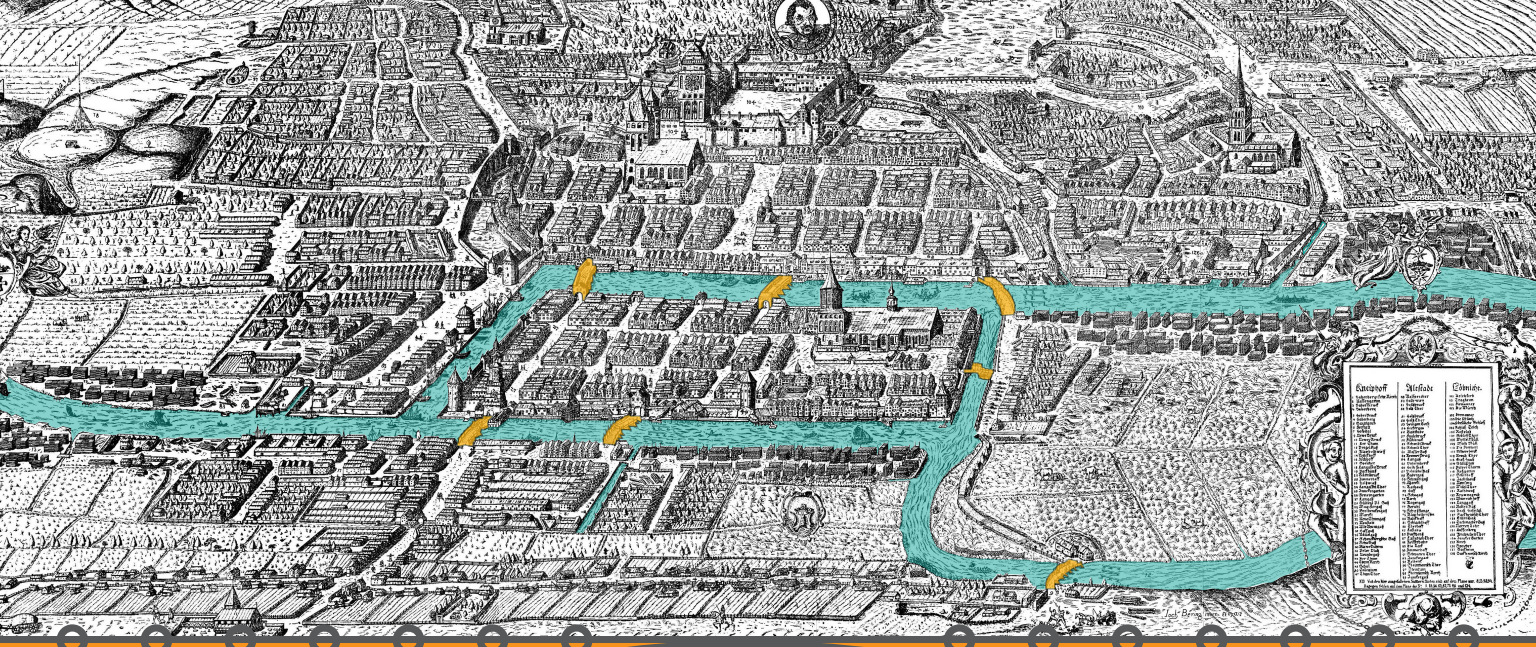
Draft Recommendations

Recommendation #1

- Establish a national structure of CI expertise/Campus Champions
  - Answer "Who do you ask about X?"
  - Vertical and horizontal social connections
  - Establish trusted parties for cross-domain
  - Exact methodology here needs more study
  - Needs buy-in from campus leadership
  - Project/technology neutrality
- Establish CI expertise in campuses by offering training to campus CI personnel.
  - Both domain and collaboration support knowledge
  - Incentives: Recognition, travel costs
- Provide mechanism for harvesting experiences, requirements, use cases, etc.
- Example: NIH Knowledge center, UK CG SIG

Recommendation #2(?)

- Establish a lowest common denominator of services
  - E.g. NSF Net and TCP/IP
- Data management plan requirement could push for campus archival
- Identity/access control seems to be underway. What's next?
  - Several things discussed: name spaces, DBs, file systems, allocations policies.
- For CI to succeed, there needs a driving vision, principles, use cases, blueprint/architecture across the directorates.
  - Independently designed as to not be overly complex for complexity/funding sake
  - There need to be a responsibility party to do this. There isn't a clear body to do this design today
  - How do we organize the vision development and architecture process to be neutral and effective?
  - Need to coordinate with other agencies.
  - Needs to be sustained.



The cover image is based on Joachim Bering's etching of the city of Königsberg, Prussia as of 1613 (now Kaliningrad, Russia). Seven bridges connect two islands in the Pregal River and the portions of the city on the bank. The mathematical problem of the Seven Bridges of Königsberg is to find a path through the city that crosses each bridge once and only once. Euler proved in 1736 that no solution to this problem exists or could exist. This image appears on the cover of each of the Campus Bridging Workshop reports.

The goal of campus bridging is to enable the seamlessly integrated use among a scientist or engineer's personal cyberinfrastructure; cyberinfrastructure on the scientist's campus; cyberinfrastructure at other campuses; and cyberinfrastructure at the regional, national, and international levels; as if they were proximate to the scientist. When working within the context of a Virtual Organization (VO), the goal of campus bridging is to make the 'virtual' aspect of the organization irrelevant (or helpful) to the work of the VO. The challenges of effective bridging of campus cyberinfrastructure are real and challenging – but not insolvable if the US open science and engineering research community works together with focus on the greater good of the US and the global community. Other materials related to campus bridging may be found at: https://pti.iu.edu/campusbridging/