SHIFTING MODALITIES OF USE IN THE XSEDE PROJECT

Richard Knepper

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics and Computing
Indiana University
May, 2017

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee:

Nathan Ensmenger, PhD

Christena Nippert-Eng, PhD

Katy Börner, PhD

Kosali Simon, PhD

April 25, 2017

For Ula, who always championed me through setbacks and successes.

For Dick, who always provided the hardest questions.

For David, who always asked "Now, how do we get you out of here?"

Acknowledgements

This dissertation is the product of a particular set of experiences in time, which were afforded to me by my concurrent role as a professional IT manager working in the XSEDE project and as a student curious about organizations, technology, and scientists. I could not have had the access, frankness, and trust that I was able to achieve without the openness and interest of the members of the XSEDE project, its staff, management, and users. I especially thank XSEDE principal investigator John Towns, who took the time to read associated papers and discuss the project with me while in the middle of running a multimillion-dollar project distributed across fifteen partner institutions, and XSEDE's NSF Program Director Rudolf Eigenmann, who provided me a view of the NSF processes behind the organization. Data for the quantitative investigations that made up part of this project were made available thanks to Dave Hart of the National Center for Atmospheric Research for providing publications information from the XSEDE portal and discussing utilization and publication information, and also to Robert Deleon and Tom Furlani of the University of Buffalo for providing XSEDE project and usage data. As an interview respondent, Dr. Craig Stewart provided excellent material as well as advice during the preparation of this research.

This project would not have been possible to bring to completion without its research committee. Dr. Nathan Ensmenger provided unwavering support and assurance throughout the process of developing the area of inquiry and the points of interest which came out of my conversations with informants. Dr. Christena Nippert-Eng was a source of guidance on matters ethnological as well as in the drafting stages of this dissertation. Dr. Katy Börner informed and instructed on the quantitative and network science portions of the research. Dr. Kosali Simon provided insight into policy matters and the policymaker perspective.

The work below would not have been possible without the guidance and support of the late David Hakken, who provided intellectual guidance and questions which provided the foundations of inquiry. David was a credit to his field and always supported those investigators who looked at the unusual, the failures, and the imbalances in our world.

Richard Knepper
Shifting Modalities of Use in the XSEDE Project

The National Science Foundation's Extreme Science and Engineering Discovery Environment (XSEDE) provides a broad range of computational resources for researchers in the United States. While the model of computational service provision that XSEDE provides suits a traditional high performance computing model well, there are some indications that computational research is moving to new modalities of use and new virtual organizations. This ethnography of the users, management, and staff of XSEDE explores how a large virtual organization is charged with providing both next-generation computational resources and reaching out to a broader community which is just beginning to embrace computational techniques. The challenges of balancing the demands of a diverse set of stakeholders, the guidance of changing leadership at the NSF, and managing the development and professionalization of the cyberinfrastructure workforce create a number of tensions and dynamics. This research project identifies how these dynamics arise and are resolved, reflecting on the science policy initiatives which define and constrain the environment. In addition to ethnographic description of the project and interview analysis, social network analyses of the XSEDE user base and research project usage information are conducted, showing a collection of traditional HPC users and support for research from traditional research oriented institutions, rather than extending to new, broader communities.

Nathan Ensmenger, PhD

Christena Nippert-Eng, PhD

Katy Börner, PhD

Kosali Simon, PhD

Contents

Dedication **iii**

Acknowledgements **iv**

Abstract **v**

1 Introduction **1**
1.1 Central Questions 6
1.2 Motivation 10
1.2.1 Informing studies of scientific collaboration 11
1.2.2 Understanding relationships between infrastructure and research 13
1.2.3 Informing science policy and NSF initiatives 18
1.3 Key Concepts 19
1.3.1 Cyberinfrastructure and the CI community 20
1.3.2 Government-sponsored cyberinfrastructure environment 32

2 Literature Review **40**
2.1 Studies of Cyberinfrastructure 40
2.1.1 The roots of cyberinfrastructure within Big Science . . 40
2.1.2 The development of collaborative cyberinfrastructure 53
2.1.3 Literature on cyberinfrastructure usage and scientific production 59
2.2 Science Policy 62
2.2.1 Government funding for basic research 64
2.2.2 Making Grants Measurable 71
2.2.3 The competitive element 75
2.2.4 Public Management Networks for Service Delivery . . 77

3 NSF-funded cyberinfrastructure prior to XSEDE 87
3.1 The NSF Supercomputing Centers Program 88
3.2 The TeraGrid 91
3.2.1 TeraGrid innovations 93
3.2.2 TeraGrid Metrics and Incentives 100
3.2.3 TeraGrid tensions and dynamics 104

4 Research Methods 112
4.1 Investigator Statement 113
4.2 Qualitative Analyses 116
4.2.1 Document Analysis 117
4.2.2 Interviews 119
4.2.3 Ethnographic Observations 125
4.3 Quantitative Analyses 129
4.3.1 Data Gathering 130
4.3.2 Bibliometric Analysis 133
4.3.3 Analysis of usage data 134
4.4 Institutional Review 135

5 Findings 136
5.1 From TeraGrid to XSEDE 137
5.1.1 The XD solicitation 137
5.1.2 The "shotgun wedding" 140
5.1.3 XSEDE operations begin: the move towards service . 144
5.2 Understanding XSEDE users 155
5.2.1 Changing fields of science in XSEDE 155
5.2.2 The "Long Tail" of XSEDE 159
5.2.3 How users leverage XSEDE 167
5.2.4 Researcher tactics 170
5.3 Understanding XSEDE and the CI community 175
5.3.1 Cyberinfrastructure Crises 178
5.3.2 Adaptations in the Virtual Organization 181

6 Conclusions 185
6.1 Cyberinfrastructure trends 186
6.2 XSEDE 2 and what comes after 189
6.3 Science Policy lessons 193
6.4 Areas for further research 195

Bibliography **197**

A Interview Questions **215**

Curriculum Vitae

Chapter 1

Introduction

Scientific research requires tools and instruments in order to conduct experiments, make measurements, and calculate results. As scientific studies have moved from individual to group efforts and technological change has moved scientific work from the laboratory to the computer, the primary instrument of inquiry has become the computer. Joint scientific work which is mediated by the Internet is increasingly the normal form of scientific inquiry [59]. With the development of these "collaboratories" as described by Finholt and Olsen [61], the work of scientists increasingly becomes more collaborative, both in terms of direct collaboration with co-authors as well as indirectly, with other researchers who develop scientific software, technicians and engineers who manage instruments and systems, administrative staff who oversee project work, and granting agency program officers who determine funding levels and manage solicitations for projects. Obtaining and managing computational resources has become a key component of a number of disciplines' research activities, and the need for computationally-skilled scientists seems to be

more urgent than ever. For some, computational methods represent a technique of scientific inquiry in addition to theory and experimentation [84]. Others dispute the computational sciences as secondary to the establishment and use of the scientific method [158]. Nevertheless, demand for computational resources continues to grow and scientists continue to produce research based upon computational methods. While the environment for research computing in the United States is still strongly tied to the original efforts to create programmable computer systems, the community of high-performance computing users is shifting and changing as new disciplines become more computationally oriented and more users with different needs and backgrounds begin to require computing resources in order to further their research goals.

Another word for the computational resources used to support research is *cyberinfrastructure*. Cyberinfrastructure is composed of computing systems, software, networking, storage, and skilled humans that maintain all of these interdependent systems [149]. Cyberinfrastructure ranges from individual laboratory computers to high-powered supercomputing systems, but like all of the infrastructures which underlie and support it, the basic utility of cyberinfrastructure comes from interconnection of components, which are created and maintained in different environments and time periods. The linking together of components is a central part of cyberinfrastructure. As early as 1971 [92], it was clear that providing scientists with access to computational resources was a neces-

sity. The next natural step was to make resources available to scientists regardless of their location. That link between researchers in disparate locations from instruments has been essential in building the fabric of the national research environment. The result has been the creation of a complex network of infrastructures available to researchers at multiple levels and scales, which offers a number of alternatives for conducting computational research. While there are patterns in developing computational resources, there do not appear to be broadly-established norms of use.

The aim of this research project is to understand the role cyberinfrastructure plays in supporting basic research and to understand how the organizations that provide cyberinfrastructure adapt to meet the needs of the scientific community. These changes are the result of a number of different activities: user choice, organizational strategies, and funding initiatives by the National Science Foundation (NSF). By understanding the interrelated activities of users, virtual organizations that support them, and funding agencies, informed by theories of Science and Technology Studies, I intend to provide insights that may be useful to the future management of cyberinfrastructure projects and to cyberinfrastucture funding initiatives put in place to further national research goals.

In order to investigate cyberinfrastructure organizations and their uses, strategies, and changes, I examine the Extreme Science and Engineering Discovery Environment (XSEDE - http://www.xsede.org), a collaborative

project funded by the National Science Foundation to provide computational resources in support of basic research [12, 154]. The XSEDE project is a 5-year, 17-partner project, funded at $121M US dollars, with responsibilities which range from providing operational support for supercomputing center activities to code optimization to user education and outreach. Awards that fund resources for Service Provider systems range up to 77M USD for the Kraken supercomputer and 55M USD for the Stampede supercomputer. At the date of this writing, continuing investments are planned for an additional 5 years of the XSEDE program. The XSEDE project represents a foundational investment in general- purpose computing in support of basic research. Understanding the processes by which XSEDE changes its service delivery offerings provides information for drafting of future science cyberinfrastructure initiatives. In a time of limited funding for resources in support of basic science, common infrastructure initiatives are one way to provide resources to a broad range of researchers. These largest of these investments are concentrated in the Supercomputing Centers, which have developed a high level of expertise in implementing and maintaining these systems, but the gains are distributed to researchers across the country [15].

I describe the NSF's funding initiatives and history of computational programs, preceding XSEDE, in section 3. I make use of both qualitative and quantitative techniques to look at the XSEDE project's activities and understand the organization, its users, and its environment. The focus,

techniques, and aims of the research were reviewed and approved by the XSEDE principal investigator and NSF program officer. While the project is funded through a grant from the NSF and therefore subject to Freedom of Information Act (FOIA), the project leadership allowed me to leverage my access to materials as part of my job responsibilities, rather than pursuing FOIA requests for information, with the goal of providing an overall open framework that supported better description and review of the project. In order to get rich descriptive information about the people involved, I conducted twenty-two interviews with XSEDE users, staff, and NSF program officers. Since the XSEDE project has extensive planning, reporting, and communication activities, I conducted document analysis over a broad swath of XSEDE literature as well as NSF publications and documents. The XSEDE Senior Management Team allowed access to a large number of internal documents and systems used to manage the processes of work to support XSEDE. With the blessings of the leadership, I also engaged in participant observation with the project's management team, as part of my professional responsibilities working in XSEDE's Campus Bridging team from 2011-2016. Participant observation activities covered management meetings, project team activities, conferences, and outreach events to campus IT staff and researchers. Quantitative components of the research project involve the analysis of researcher publications and usage of computing resources to understand what fields of science are making use of resources, and what collaborative relationships exist within the XSEDE

project.

As XSEDE provides such a broad range of services to the scientific community, the organization's leadership must make choices as to the communities and fields it serves, in order to provide benefits that meet NSF criteria for funded projects. This results in tensions between varying communities and between technology choices. Some of these tensions are related to allocation of resources, some are about the kinds of activities that XSEDE engages in and supports, some of them are about the NSF's strategies for establishing new resources. This dissertation research project focuses on the factors, such as the relationship between parter organizations, the requirements of the NSF to meet a broad range of community needs, and the changing backdrop of computational techniques, which create these tensions.

1.1 Central Questions

The central questions of this research focus on the changing needs of researchers in the U.S. for resources, and the NSF's initiatives to create cyberinfrastructure resources which address the distribution of computational resources across the range of academic institutions, from high-powered research institutions to educational institutions and workforce development programs. While the world of personal computing changes constantly, use of computers for research has largely remained the same as it was since the development of the teletype machine and batch-processing

of computational jobs. Researchers interact with a terminal, the modern equivalent of a teletype machine, to create a batch job submission, wait for their program instructions to go through a processing queue, and then interpret the results from their submission [54]. In the meantime, in our daily lives, we use computers with graphical interfaces, by touch, and by speech in order to accomplish any number of tasks. New generations of users are entering the scientific workforce who are likely more familiar with these forms of personal computing which have become commonplace than with the command-line environment. Furthermore, new fields of science are adopting computational techniques which further their ability to ask questions and develop theories, and the analyses these fields rely on have a different way of using computing power than the traditional big users of computers: physicists, chemists, and engineers [30, 165].

Who is XSEDE's user base? What are their needs? How do they get what they need from the organization? If XSEDE is a general-purpose cyberinfrastructure, then it is important for us to understand for whom this cyberinfrastructure is built and extended. The intended user-base for XSEDE consists of users who have intensive computational needs and can make use of large-scale parallel systems; but also users who have modest needs and no access to such systems at their home institution. XSEDE balances these needs by providing systems that can process both large workloads and integrate many small ones. However, technology does not impact all equally, different individuals benefit or bear costs

due to the varying impacts of technology [102, 166], and XSEDE as a cyberinfrastructure virtual organization is no different. Indeed, as this study demonstrates, the distributed configuration of XSEDE as a virtual organization means that varying groups within the organization have different missions, goals, and agendas to pursue. Furthermore, XSEDE's users exhibit strategic behavior in acquiring resources, linking activities, leveraging the importance of the project, and building legitimacy across their own initiatives. Based on the changing use of technology over time, a number of the informants interviewed for this research expressed that demand for the computational services provided by XSEDE are in the process of changing, with the distribution of needs changing shape over time as different fields of science adopt XSEDE in greater numbers [68]. By asking these questions I intend to understand the process of this change in user base, and what kind of benefits users are able to extract from a project with very broad cope of mission and many metrics to meet.

How do organizations like XSEDE adapt to meet needs? XSEDE is engaged in activities to meet the need for basic cyberinfrastructure, and those needs represent a moving target. Traditionally consumers of high-performance computational capabilities have been fields in the physical sciences which are producing computational work at significant scale. While the demand for computational infrastructure for traditional models of high-performance computing continues to grow as these disciplines undertake ever larger and more detailed analyses [54, 148], thanks to

the ability to capture and analyze data at large scale, fields which previously had minimal computational needs have changed the kind of questions asked and the methods of finding answers to make use of computational techniques which can inform their field. As Hallam Stevens notes for the biological sciences, during the 1990's, computational techniques, largely viewed as unnecessary by biologists for a number of years, were paired with sequencing methods which generated a large amount of data tractable to statistical analysis, allowing for questions to become broader, change focus from individual processes to global processes, and develop comparative analyses [146].

What should the NSF pursue in its cyberinfrastructure initiatives to meet the needs of the scientific community? The NSF develops policies for support of research activities. Its development of cyberinfrastructure for support of basic research involves two types of initiatives. The Division of Computing and Communications Foundations (CCF) and Division of Computer and Network Systems (CNS) manage computer science and engineering initiatives to create next-generation resources that can support research with extreme needs for computational capacity, data, and responsiveness. In the Office for Advanced Cyberinfrastructure (OCI), initiatives provide the general research community with additional resources which can support a broad range of computational needs [9]. Individual instrumentation awards can support acquisitions of high-performance systems which can be shared with the broader research community. The

responsibility of the NSF is to provide a comprehensive program which supports the development of computer science and engineering while also providing capability computing in order to facilitate the research of scientists who do not have sufficient capability at their home institutions. To that end, the XSEDE project and associated NSF-funded initiatives are initiatives which knit together the cyberinfrastructure resources in aggregate.

1.2 Motivation

This research project is motivated on the principal of improving our undestanding of how scientists collaborate and interact, how cyberinfrastructure supports research activities, and what policy initiatives best suit general purpose cyberinfrastructure support. The basis of this inquiry revolves around the habits and behaviors of scientists, how they cooperate with each other and acquire resources in order to further their agendas. Furthermore, this is intended to be an exploration of the relationship between infrastructure, specifically cyberinfrastructure, and scientific work. While all science makes use of instruments and equipment which have their own frailties [54], I argue that most implementations of cyberinfrastructure are either research projects in their own rights, "production" systems which support research, or some combination of the two. And finally, I hope to make use of the conclusions here in order to help the XSEDE project leadership continue to develop their own project in order

to more effectively meet the needs of scientists, and to provide input to the National Science Foundation on its initiatives to fund cyberinfrastructure in general.

1.2.1 Informing studies of scientific collaboration

By examining the changing use of XSEDE by different fields of science and different researchers, I hope to gather understanding about the larger relationship of computational effort to scientific endeavors. A number of scholars of science and technology studies examine the relationship between infrastructure and science as more and more research projects grow in scale and number of collaborators [94, 70, 101, 54, 170]. While cyberinfrastructure shares many of the features of collaborative science research: they are large projects, with many staff distributed across a range of organizations, they are managed based on production of scholarly work; cyberinfrastructure is different in that it is the service delivery activity for scientific projects, rather than being a project of inquiry in and of itself. As mentioned above, the notion that computational techniques represent a new development of scientific inquiry is a disputed one. A number of disciplines are adopting computational techniques for discovery for the first time, with new possibilities for inquiry made possible by simulation, modeling, and statistical inference [165, **?**] The advocates of computational perspectives note that new techniques for analysis represent potential for new results and new conceptual views of prior problems.

Others have questions about the theoretical underpinnings of computational approaches. Understanding the relationship between research and computational activities informs us about the possible conclusions that research can make, and understanding the shift in approaches provides perspective on the direction of inquiry. By broadening the diversity of its domains and users, the NSF is attempting to bring new theories and new perspectives to bear on problems, but I argue that this change is not a one-way street: by offering resources to these users and disciplines, the inquiries of these users and disciplines are also affected in terms of the questions asked and the means of arriving at answers.

Investigating the XSEDE project also provides opportunities for informing the study of virtual organizations. Virtual organizations are those which decouple management from service delivery, are frequently decentralized, and whose members come from a "home organization" which may have another function than the virtual organization [111, 64]. Features of the XSEDE organization, namely a distributed management structure and tiered service provider system, allow for exploration of the project as an exemplar virtual organization, which employs significant efforts to address dual-loyalty issues and resource constraints. XSEDE deals with some considerable interorganizational incentives, as many of the partner organizations are competitive rivals for NSF funding of other projects, including large awards for Service Provider systems. In the area of resource constraints, XSEDE, while a highly-funded program in comparison

to many NSF awards, is also atypical in its mission and scale, and stands out from other projects as a result of both of these characteristics, making it a target for questions about the appropriateness of the project mission and funding levels. As a virtual organization, XSEDE also deals with issues of geographical distribution and significantly leverages technology in order to allow a management team at 15 different physical locations to work together effectively. Some of my findings provide insight into how XSEDE works to manage distance and time differences in order to have a fully-integrated management team.

1.2.2 Understanding relationships between infrastructure and research

Scientific inquiry is empirical in nature and relies on the use of common tools to provide measurements that can be viewed and inspected by other scientists [54, 139, 101]. Instruments are the bedrock of scientific activity, but modern instrumentation is worlds removed from the basic instruments we learn in grade school. Particle detectors and interferometers are complex systems in their own right, and can have significant differences in activity, based on the methods of extracting measurements. Even relatively simple instruments such as thermometers and barometers, when used en masse, represent a problem in standardization of measurements and collecting activities [54]. Not only do the instruments affect the resulting science, the representations and descriptions made by scientists are

intermediaries in the process of communicating between phenomena and scientific thought, and have varied over time as thought about the role of science has changed [50]. Computational infrastructure represents a different way of interacting with the practice of science. To understand how this may affect inquiry, it is useful to look at the question of reproducibility of computational research.

For research that analyzes collected data through complex computational techniques, it is a reasonable expectation to be able to replicate the algorithms used to analyze the data on another computer and verify the initial results for the data that was initially collected, if not for independently-collected data. In practice, this is a much more complex process. Even if two scientists might agree on the mathematical transformations of the data, the implementation of these transformations in code can have large effects on the resulting answers. Other factors, such as the scale of computational power and availability of data represent similar challenges to performing replication. Likewise, simulations which model systems computationally may produce more data in output than is actually tractable to manage in multiple systems, representing its own computational problem in comparing two sets of output for reproducibility. Furthermore, not all of these activities are cleanly self-contained. When a scientist makes use of multiple different software programs in order to conduct a complex analysis, it may be possible to capture individual codes, but the combination is not possible to save, except by a remarkably con-

scientious researcher, making note of her activities [122, 51]. There are some tools designed to capture workflows and make them reproducible, such as the Sci2 tool, but not all types of analysis lend themselves easily to descriptive workflows. Some authors have proposed making use of cloud technologies and system images that can be run in virtual environments to provide reproducibility [83] or by creating repositories where codes could be stored and shared for later reproducibility efforts [109]. Other authors have suggested that exact replication is a futile goal and suggest a reframing of the definition of reproducibility entirely [48].

These issues around reproducibility suggest that computational strategies employed for research have influence over the conclusions at which scientific efforts arrive. Computational techniques, which are commonly understood to be standardized – after all, the same code should provide the same results when given the same data – are actually variable and dependent on other factors. The computational world, instead of being more certain than the physical world, has instruments and methods that are just as fraught with issues as the laboratory. One of the researchers I spoke with who uses both the Open Science Grid and XSEDE resources stated that he has begun to regard job runs as having an "experimental yield", in that not all runs will succeed, and not all runs will return the expected results. Results of analysis depend on the complex interaction of code, data, and systems which the code runs on.

The matter is further complicated when cyberinfrastructure resources

are experiments in their own right. Large systems, such as the NSF-funded Blue Waters and Department of Energy systems, have, by merit of their size and complexity, an experimental nature. The Blue Waters system is a 13 petaFLOP system at University of Illinois Urbana-Champaign designed to support research that requires the highest level of computational capacity. Blue Waters provides resources for highly complex computational research and also serves as a platform for exploring the issues around supporting computation at the petaFLOP scale. During its development and implementation, which would create the largest computer cluster ever built by the NSF, the project was forced to change vendors due to difficulties with implementing the original vendor IBM's design [57], and once implemented in partnership with Cray, provided considerable understanding of problems of concurrency and failure state in large computer cluster systems [**?**]. The Department of Energy in their own right have implemented a significant number of these large scale systems at their "Leadership Computing Facilities" at Oak Ridge National Labs, Argonne National Labs, and the National Energy Research Scientific Computing Center. These systems currently range up to the 27-petaFLOP TITAN system at Oak Ridge, with planned upgrades that will provide a 200-petaFLOP system designated Summit by the end of 2018 [4].

These systems are at the edge of attainable computing scale and managing parallel execution, not to mention component failures, are an area of investigation in computer science in its own right. A 2009 study of so-

called “Extreme Scale” systems identified that the number of systems in a cluster is not the only issue: the increasing density of systems, which allows a petaFLOPS sytem to be located in a departmental lab or a single computer to achieve teraFLOPS performance, is the root of problems in concurrency, energy efficiency, and resiliency [21]. Concurrency refers to the need for thousands of processors to be able to work in concert with each other and communicate with each other across the entire system. Energy efficiency refers to the ability of system components to deliver significant increases in performance without similar increases in energy consumption and attendant heat production. Resiliency is defined as the ability of a system to keep running despite the failure of individual components. Building cyberinfrastructure requires the successful incorporation of a broad set of disciplines working together to create a highly complex sytstem. Engineering a bridge requires principles from civil and mechanical engineering. Development of a power distribution grid requires electrical engineering, transmission and conditioning expertise, and knowledge about levels of utilization. In contrast to basic infrastructure which may rely on a few disciplines to create a system successfully, cyberinfrastructure requires engineering principles from the silicon of the chips involved, to power and cooling issues, networking infrastructure, storage, and software engineering. Thus, the development and extension of cyberinfrastructure is a highly complex process, with questions at multiple levels which determine the outcomes of the analyses conducted.

1.2.3 Informing science policy and NSF initiatives

The final goal of this investigation is to build understanding that will inform the management of XSEDE about how change occurs within the project and the initiatives which follow it. The field of study of Science, Technology and Society (STS), particularly the idea of "cycles of credit" which are proferred by Latour and Woolgar, has a significant amount of resonance with the activities of XSEDE project, particularly the allocation of resources, itself a sort of microcosm of the NSF funding environment. Previous research has looked at the predecessor project to XSEDE, the TeraGrid, and made contributions both to the structuring of the solicitations which led to XSEDE and to the management choices within XSEDE [41, 172, 173]. Because XSEDE is a project in part aimed at addressing distributional issues by making resources available to those who might not otherwise have access, with the understanding that the process of evaluating science and allocating resources is a complex one. My research project is aimed at providing a translational function that utilizes applicable STS theory in order to inform policy choices by XSEDE and the NSF.

There is reason to believe that XSEDE will act on these recommendations. The XSEDE project as a rule is intensely self-scrutinizing, and has made a number of changes in response to feedback from its own external advisory board, the NSF, and internal evaluations. XSEDE must navigate the tension between providing a next-generation computing platform and providing one that supports discovery within the "long tail" of

science. The technical expertise of XSEDE and Service Provider staff supports providing an environment that is suitable for both types of users. Understanding how to select, implement, and disseminate technologies that leverage this considerable computational and technical capacity will serve the organization well during its next 5 year term.

Furthermore, the NSF is closely examining the progress of XSEDE as well as that of the Open Science Grid in order to evaluate the future of cyberinfrastructure investment at the national level [15]. The NSF investment in XSEDE is atypical for the organization, which normally funds individual research initiatives based on basic research projects. As the NSF continues to deal with a political environment that focuses on the reduction of funding for scientific programs, it is critical that programs deliver their intended results as efficiently as they do effectively. Part of the motiviation of pursuing this research is to provide reflections that are useful to the NSF on drafting programs and solicitations for its ongoing Computer and Information Science and Engineering Advanced Cyberinfrastructure initiatives.

1.3 Key Concepts

In this section I illustrate some of the key concepts and standard nomenclature surrounding the cyberinfrastructure community in the US, in order to make later descriptions of the XSEDE project and its workings more clear. This section discusses the environment of government-sponsored

cyberinfrastructure as well as common concepts of cyberinfrastructure. This section is provided with the intention to give some basis of understanding of the complex environment in which my informants are immersed, in which scientists work in university and government projects to pursue basic research, a mix of computer science, engineering, science policy, and academic cycles of credit, in which there is significant debate about how best to provide the infrastructure which supports basic science. My hope is that by describing the national cyberinfrastructure environment well, I can provide sufficient thick description [73] in later sections without having to provide considerable supporting exposition, and furthermore to be able to encapsulate the description of cyberinfrastructure to be used in later investigations. Below I discuss some of the physical components of US cyberinfrastructure as well as its underlying policy initiatives.

1.3.1 Cyberinfrastructure and the CI community

As discussed earlier, cyberinfrastructure is the aggregate of computational, storage, networking software, and human resources which supports research activities. While it is common to think of a brightly-lit supercomputer in a datacenter as central to cyberinfrastructure, this represents only one element of an interconnected system with distributed parts under the control of different individuals and organizations. The *cyberinfrastructure community*, then, is the community of provders and users of

cyberinfrastructure, including supercomputing center and university IT staff, scientific software developers (funded by research grants or by operating budgets), researchers and graduate students making use of the computing systems, and policy-makers, such as NSF program officers. The CI community is also influenced by other visible figures: leaders of large projects (such as XSEDE and the Open Science Grid, but also others), directors and executives of supercomputing centers. This environment makes for a mix that would be easy to mistake for an example of the "garbage can theory" of organizational choice [46]. Attending a conference such as the International Conference for High Performance Computing, Networking, Storage, and Analysis (commonly referred to simply as "SC", for Supercomputing), one sees a dazzling array of choices: hardware, software, organizations, identities, and sectors. The CI community, however, has its own norms and standards and means for making choices. Rather than being a true garbage can, consisting of a mix of solutions, issues, and actors, where the choices that connect them are determined by happenstance and proximity, there are processes for selecting and using different solutions, which often correspond to the credibility of a given solution more readily than the effectiveness of its alternatives. This leads to a kind of inertia among many actors in cyberinfrastructure, which seems to have affected the development of solutions over the long term.

Tracing cyberinfrastructure from the bottom up as other scholars of science and computing [54, 101] have done, the most common means of

making use of cyberinfrastructure would be the laptop computer. This is the common, base unit of computation for most individuals involved in scientific research. The laptop computer provides facilities for recording data, performing basic analyses, and writing up results in a portable package that can be easily transported from the researcher's office to classrooms, laboratories, and conferences. Scientists, students, technicians, and administrators commonly interact with this resource all day, every day. For some scientists, an office or lab workstation is required to interface with analog equipment, or complete calculations that are feasible on a larger system but not a laptop. From the laptop or desktop system, users are connected by networks, created and maintained by institutions and businesses, to other systems that we make use of every day. In the computational context, these systems can connect to the supercomputer in the datacenter just mentioned, but they can also make use of so-called high throughput systems, or science gateway systems. Most recently the "compute condo" and "research cloud" models have become more prevalent, drawing examples from private IT service and cloud providers. Below I detail some of the common modalities of computational science from the point of view of cyberinfrastructure providers – those administrative staff at universities and research centers charged with building, maintaining, and supporting computational resources.

High Performance Computing

High performance computing (HPC) is the embodiment of the supercomputing system as we commonly imagine it, but there are any number of manifestations of this computational modality. High performance computing frequently refers specifically to computing that requires parallel processing, analyses that require more than a single processor or machine to complete. While there are a multitude of implementations of HPC systems, from a small cluster of machines in a lab environment to the large-scale implementations at Department of Energy datacenters, there are a few characteristics which determine the type of use within a high performance computing environment and the requirements of systems that can be classified as HPC systems. HPC systems are characterized by the use of similar or identical systems working together, joined by an interconnect of some sort, which have a scheduling system that allows for the submission of jobs and manages allocation of work across the system. Early high performance systems, an outgrowth of the Unix systems used for large scale automation, business, and database implementations used shared memory across a large number of processors, but the expense of this design and the arrival of the open source Linux operating system brought forth the commodity cluster system. Commodity clusters provide parallel processing capability based on common server hardware, making resources more affordable and expandable.

Even with the efficiencies provided by commoditization of hardware,

large-scale high performance systems can represent significant capital expenditures for campuses. Particularly in an era of reduced funding for research infrastructure, acquiring resources can be difficult for universities which want to provide modest local compute resources. A number of universities have engaged in initiatives to buy "community condominium" clusters. These "condo" systems are generally acquired in stages. Campus cyberinfrastructure units build a base cluster resource and scientists at the university are able to purchase additional systems that can be added to the cluster and managed as part of the overall system. Generally, within the condo model, the purchasing scientists have preferential access to the system provided by the scheduling system and are able to submit jobs at any time. Meanwhile, the cluster is administrated by computing center staff, and is located in a secure location, rather than in a departmental lab, office, or storage closet. Campuses which provide condo clusters typically work with faculty with start-up funds or grants to purchase systems, based on scientists' needs. Advocates of the condo model note that this allows administration and security responsibilities to fall on IT staff rather than graduate students, and provides researchers with the computational resources needed, with the added benefit that system lifecycle can be better maintained monitored and regulated in the campus datacenter [23, 19, 33].

HPC system users access these resources much as they have since the 1980's, through a terminal emulation program (itself a software replace-

ment for a teletype terminal) which allows them to enter commands and interact with text interfaces. Through this fairly basic interface, users manage data, create job scripts, and submit jobs. Jobs are processed by a scheduling system which allocates work across the supercomputer in order to ensure that the largest portion of the system is working to meet demand. The scheduling system manages starting jobs on the individual computers that make up the computer system, delivering data, and returning results to the scientist.

HPC systems are designed with parallel processing workloads in mind, but in common use, these resources may be used a number of ways, from system-wide parameter sweeps and large-scale simulations that require a large percentage of the resource to be working together and communicating with each other, to other modalities that are either not parallel approaches to problems ('serial jobs' which run on a single computer), or they require input from the user and cannot be run via a scheduler system ('interactive jobs'). Modern supercomputing centers tend to provide resources for all three types of jobs, although they may provide different resources for each type in an effort to have the most effective allocation of resources for all of their users. Jobs at university supercomputing centers which utilize the system in its entirety are fairly rare, so the scheduling system is responsible for "packing" jobs into the system in an efficient way that allows more jobs to run concurrently. Centers that deliver a large number of completed jobs, then, with system utilization at a level

of 80% or more, are delivering service efficiently and effectively, as a supercomputer system that is even half idle is a waste of power and cooling resources. If the basic unit of service delivery for an HPC system, then, is a job, then the atomic metric which allows one job to be compared to another as well as capturing work delivered over a period of time, is the core-hour. The core- hour, or the work produced by one computer processor core during one hour's time, and many centers and projects provide statistics on core-hours to their stakeholders in order to show the amount of computational work delivered, what percent of time was spent in planned or unplanned downtime, and what types of usage are most prevalent on systems.

The level of interprocess or inter-machine communication that needs to take place determines the extent to which a particular code is parallel. *Tightly-coupled* codes require fast networks in between systems in order to reduce latency between the calculation states of the codes being executed. Depending on the needs of the particular analysis and the level of concurrency involved, the network interconnect can be exceedingly important to the performance of a given code on a given system. Highly parallel codes, referred to as "pleasingly" or sometimes "embarrassingly" parallel, can be executed on a number of systems at once with less reliance on interprocess communication. These highly parallel codes can be run on systems with higher latency interconnects. The looser coupling of analyses involved in these types of problems has given rise to a more op-

portunistic way of acquiring computational resources, commonly referred to as high throughput computing (HTC).

High Throughput Computing

For projects that are based in highly parallel codes, such as the popular "volunteer science" software Folding@Home or BOINC, these projects work by distributing highly parallel calculations that are run on millions of personal computers when they are not being used, with an architecture provided by the project for job assignment, data collection, and statistics and reporting [22, 24]. The broad availability of computational resources that are not otherwise available has given rise to opportunistic use of computer resources, such as computer labs or personal computers in offices, but also to opportunistic use of larger resources. The Open Science Grid makes extensive use of computer resources both small and large to run jobs for a number of large-scale collaborative scientific projects, most notably the analysis of data from the Large Hadron Collider project at CERN and the SB Grid structural biology project, but also supporting 92 different research projects, including materials science, bacteriology, neuroscience, and economics, with grid computing resources that are largely volunteered from the scientific community [126, 118].

In contrast to HPC implementations which largely concentrate on dense architectures of systems with close interconnects, HTC implementations are less dependent on these types of resources, although they can and do

take advantage of volunteer HPC systems in addition to other resources. The Open Science Grid delivers considerable capability to its dozens member Virtual Organizations: about 800 million core-hours per year as of 2015 from systems volunteered by its member VO's and other sources, such as computer systems and clusters on university campuses. The modality of use for HTC is much the same as for HPC: users access systems via a terminal, or submit their jobs via terminal through their own system that is attached to the computational grid.

Science gateways

Yet another common modality of use for scientific computation is the science gateway. Rather than forcing users to interact with systems via terminal, science gateways provide a web-based functionality for usage. The web interface, through a set of software known as middleware that works between the web server and HPC or HTC resources, provides facilities where scientists can choose software for analysis, manage data, and track the provenance of computational experiments. The middleware services manage authentication, flows of data, invocation and execution of applications on HPC resources, as well as monitoring services which report back to the user [124]. Computational execution is handled by HPC or HTC resources, meaning that the science gateway does not provide actual computational services, rather it provides management of these activities that would otherwise be directly coordinated and completed by the user.

This approach means that it is possible to manage some optimization of tasks that would otherwise have to be discovered by the individual scientist, and such optimizations can be applied for every user of the gateway.

Science gateways are commonly built for communities of scientists, either around a particular domain, such as the SEAGrid and nanoHUB gateways, or they can be ways of making university cyberinfrastructure available to faculty at an institution. Science gateways are also a means for gathering many streams of research under one website, including citizen science initiatives or multidisciplinary work [136]. The science gateway approach allows for community members and in some cases lay persons to have a single web site under which data, analyses, and visualizations can be captured and shared, results can be propagated, and members can work collaboratively to develop scientific findings. Gateways can also provide resources for training and teaching concepts of computational analysis while allowing middleware to reduce the technical overhead required for getting started.

Research clouds

With the rise of cloud-based computing initiatives such as Amazon EC2, Microsoft Azure, Google App Engine, and others, researchers have frequently looked at cloud-based computational capabilities for conducting work. These offerings are frequently divided into "private clouds" in which the tenant has sole access to hardware dedicated to that user or com-

pany; “public clouds” on the other hand, have multiple tenants who are provided a part of a common environment; an “on-premise cloud” provides dedicated hardware for a cloud that is completely owned and managed by the institution. Researcher adoption of cloud computational services has been highly varying. On the one hand, researchers are finding that it is possible to secure computer cycles with a minimal amount of grant funding, on the other, the nature of cloud use and pricing means that an absent-minded collaborator or grad student can incur significant charges. Estimating the capability of resources and time has also proved to be a problem, leaving scientists scrambling to find funds if analyses run out of allotted time before they are completed and a paper deadline is coming up. Public or private clouds, while offering computational capability on-demand at a per-hour price, provide largely atomic units of service, leaving coordination of resources up to the purchaser. Furthermore, while private clouds provide connectivity at the gigabit level, there are few offerings that can provide low-latency networks of the type that handle HPC workloads effectively. “Research clouds” take a number of forms and attempt to address some of these issues of demand for elastic resources for computational research.

The National Institutes of Health encourage extensive use of cloud resources for research. The NSF has funded a number of research cloud initiatives for investigating the use of cloud infrastructure for supporting research, with a significant amount of activity dedicated to the develop-

ment of infrastructure for instantiating and managing cloud resources for a research cloud [65, 103, 130]. More recently, the NSF has funded a "production research cloud" which dedicates a significant amount of resources to the XSEDE project. Within this research cloud, researchers can provision and start individual systems for their work, but they can also manage large numbers of systems which work together to support a cluster with parallel processing capability, science gateway systems, or other computational systems as needed. When the work is done, the system images can be archived and shared via a document object identifier (DOI) [147].

The use of research cloud systems has resulted in a number of initiatives supporting new models of using computational resources. Due to the novelty of these types of resources, computational work which can leverage a flexible backend tends to adopt more easily than new usage of research clouds. Many of the projects making extensive use of research clouds are creating and implementing back end resources for existing science gateways, as elastic resources which can be leveraged as demand varies. Coordination of research cloud resources to be elastic computational resources for new research requires extensive work to develop and improve utilization, including new middleware tools which manage cloud systems. As cloud systems become more standardized (and begin to adopt frameworks from industry), management of the resources in the cloud becomes easier to do.

All of the types of usage described here, from traditional supercomputers to research clouds, has been part of an effort to provide utility computing, described by Foster and Kesselman at the very beginning of the NSF Centers program [64]. The vision behind a great number of the projects and activities in this arena is to create a computational environment that provides the appropriate amount of resources to researchers, who can make use of systems with little to no transition cost for moving between resources.

1.3.2 Government-sponsored cyberinfrastructure environment

As of the date of this writing, the XSEDE Service Provider community provides nearly solely HPC systems for use. All of the computational resources provided by XSEDE provide an HPC capacity in some form or another (e.g. on the Jetstream research cloud system, "virtual" clusters can be created from virtual machines running in the cloud). XSEDE Service Providers provide around 2.3 billion core-hours per year (as of 2015) from the 12 Service Provider systems currently available [118].

While the XSEDE project provides significant resources to the scientific community, there are a number of other organizations which provide HPC capabilities. This section provides detail about the network of CI organizations which provides computational resources to the research community. The environment for cyberinfrastructure investments is described in a 1993 report of of a blue-ribbon NSF commission on the "Desktop to the

TeraFLOP" [35]. In the report, a figure dubbed the "Branscomb Pyramid" describes an enviroment with a small number of HPC systems which are capable of supporting the largest scale of computational analyses at the top, supercomputers provided at centers as the next most prevalent systems, followed by smaller campus or departmental-scale systems, and finally individual workstations. This figure is shown below from the original report in Figure 1.1. The TeraFLOP scale of the original Branscomb Pyramid has been replaced with a more generalized one describing "leadership-class" systems at the very top, followed by large scale resources and center supercomputers, campus and commercial clusters, with individual computers at the base. The NSF funds initiatives at all of the levels described here, over the course of multiple strategic initiatives [14, 15]. The Cyberinfrastructure for 21st Century strategy laid out the reasoning behind NSF investments in what it termed "Track 1" and "Track 2" systems, that is, leadership and large-scale system funding initiatives.

Leadership-class systems in the United States are generally funded by awards from the Department of Energy to computing centers in the national laboratories system, which selects sites and architectures specifically to provide large-scale performance for materials science, combustion, high-energy physics, and climate modeling applications [104]. These systems are available to researchers working for the Department of Energy or in collaborative relationships. NSF has also funded a leadership-class system especially dedicated to large-scale analyses [85, 53], the Blue Wa-

Figure 1.1: The original "Branscomb Pyramid" figures, describing the US computational infrastructure

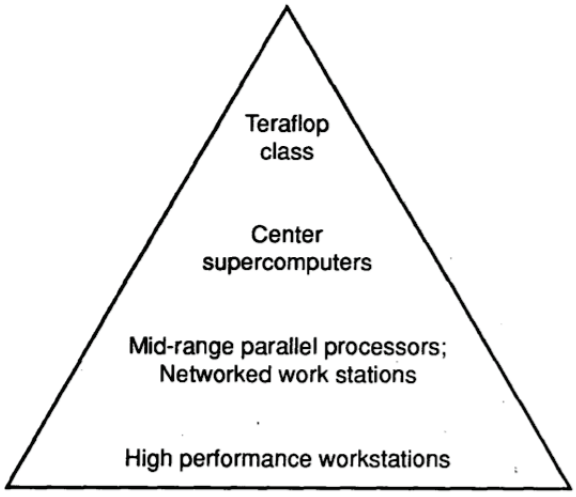


Figure A
PYRAMID OF HIGH PERFORMANCE COMPUTING ENVIRONMENTS

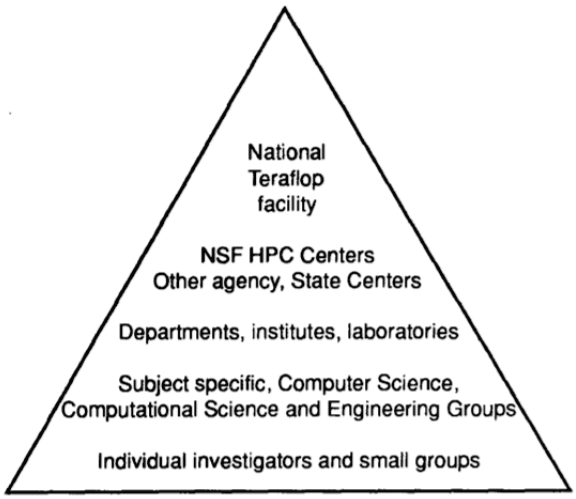


Figure B
PYRAMID OF HIGH PERFORMANCE COMPUTING INSTITUTIONS

ters system at University of Illinois Urbana-Champaign, under the NSF "Track 1" award [8, 5]. Leadership level systems at national labs and centers include Los Alamos National Laboratory's Trinity, the Cheyenne system at the National Center for Atmospheric Research, the Titan supercomputer at Oak Ridge National Lab, and others [152].

The US Department of energy operates 17 national laboratories that all provide computational capability to high-energy physics researchers, with a special program for "high-end computing facilities" at Argonne, Lawrence Berkeley, and Oak Ridge National Labs. These systems are also refered to as "leadership-class" supercomputers, which provide the largest systems available to the US scientific community. These systems are high-capability systems, able to handle extremely complex and data-intensive analyses. Allocations and rules for these systems tend to be relatively strict due to the large user population and sometimes sensitive nature of analyses. High-end computing facilities such as these have attendant storage systems which can provide large amounts of storage space with speeds sufficient to meet the needs of computations that demand fast ingestion of data, or simulations which rapidly produce data on their own. These systems are sizeable enough and complex enough that they may require their own facilities or significant modifications to existing datacenter facilities, staff to support systems and applications. The Blue Waters system required construction of its own facility, the National Petascale Computing Facility. Each of the leadership-class systems repre-

sents a significant investment in computational capability and marshalls significant resources at the center which manages it. Furthermore, implementation of systems at this scale require partnership between vendors and center staff, including the design and implementation of new processor, accellerator and networking technologies, which can be complex [57].

The following tier, "large scale resources", is made up of systems which are more broadly available, such as XSEDE systems and cluster resources available via OSG. These systems are funded through NSF awards, and most of the XSEDE systems are funded through the "Track 2" program [5], or by NASA and other agencies, which have ties to disciplines with specific computational needs. Other large scale resources in this part of the pyramid include the top end of systems funded via the NSF Major Research Instrumentation (MRI) program, which funds acquisition and development of systems between $100,000 and $4,000,000. While research instrumentation is perhaps most broadly understood as devices capable of conducting scientific observations, among the spectrometers, interferometers, telescopes, and electron microscopes, ten of the most recent fifty MRI awards at the time of this writing are for computational systems of some sort or another [7]. These center and campus resources represent considerable computational capacity for their institutions, and are in many cases used as resources for collaborative projects. The NSF Track 2 systems are awarded based on the expectation that they will be-

come XSEDE *Service Providers*, systems for physics research generally have some component dedicated to the OSG, and campus clusters are frequently used for collaborative research efforts which support scientists at different universities. Like the leadership-class systems above, these systems often have designated storage systems which provide sufficient data input and output speeds to match the requirements of the computational systems. This tier of large scale resources has a much more varied set of policies for access, based on the policies of the institution they are established at and the scope of their research mission. Large scale resources such as the NSF Supercomputing Centers systems require significant investment in staff and facilities, sometimes requiring modification of existing datacenter facilities to implement. These systems, especially those at the high end of this section of the Branscomb Pyramid, also require significant vendor input to complete implementation, but systems are generally extensions or assemblages of existing technologies.

The "medium scale" set of campus clusters described in the next tier of the pyramid describe systems which are modest campus resources or laboratory resources, at the lower end of the NSF MRI award range described above, funded as capital expenditures for other intiatives, or purchased by campuses for researcher use as general-purpose computational resources. Medium scale systems might also be negotiated for a researcher as part of a start-up package, depending on the resources involved, and the particular requirements of the field. These systems tend to be cluster

computers which may be as small as a few integrated systems deployed in the corner of a lab, or they may be modest data center resources. They may also be purpose-built systems dedicated to a particular type of analysis at a scale large enough to handle a lab or group of faculty members. Storage for these systems tends to be either integrated directly with the computers that make it up, or it may be a small networked file system resource which is provided to all of the systems in the cluster. These systems may have some staff dedicated to their management, but some smaller systems might have as little as one or two IT staff managing hardware, supporting applications, and helping researchers make use of the system.

The "base of the pyramid" consists of personal systems, those laptops mentioned above as part of every researcher's everyday activities, as well as single-computer workstations, laboratory desktop systems, and the like. These systems are capable of running scientific codes but are not able to handle either compute-intensive or data-intensive work, they are self-contained systems that can be purchased off the shelf and while they may be made of the best components available, have difficulty scaling to researcher needs. These systems are generally maintained directly by faculty members responsible for them, or perhaps by a knowledgeable graduate student, which tends to cause some issues with continuity. Individual faculty members tend to hang on to systems purchased with their own funds or funds awarded to them by the university or grants, and these

systems tend to represent security risks in the long term as graduate students move on and faculty members focus on research rather than information technology. However, these systems and their users represent a critical part of the CI environment, as they provide the means for new users to get acquainted and learn how to make use of computational resources to conduct analyses and get increasingly better answers. These systems also make up the training ground for many of the professional staff described above who maintain and develop larger scale systems. For some researchers, this foundational range is the only size of computational resource they require, and there is evidence that impactful science occurs all across the Branscomb Pyramid [10].

Chapter 2

Literature Review

2.1 Studies of Cyberinfrastructure

2.1.1 The roots of cyberinfrastructure within Big Science

Cyberinfrastructures such as XSEDE are intimately enmeshed with the development of "Big Science" projects. These projects, first described by Alvin Weinberg [162], are "large-scale, monumental enterprises" which exemplify the 20th century fascination with technology and scientific advancement. Not only do these projects (Weinberg names space travel and high-energy physics specifically) investigate basic scientific questions, they represent national prestige and cultural touchstones much the same way that architectural and artistic endeavors on a grand scale do. The American fascination with large-scale accomplishments is well-documented. Nye [117] closely examines the development and testing of the atomic bomb and the Apollo XI program as extensions of American fascination with man-made creations with scale and grandeur, demonstrative of American mastery over the environment and scientific knowledge, these partic-

ular developments –accompanied as they were with images of destruction— were also firmly linked with questions about the morality of scientific exploration, vulnerability to long-distance violence, and satellite surveillance. Certainly the impetus towards Big Science projects was increased during the Second World War, as rigorous, mission-driven project management procedures and large-scale undertakings were incorporated into the research and development community in the major engaged countries, although the direction of research endeavors towards collaborative work was already well underway by that point [70].

The proliferation of scientists and scientific work is well-documented, along with the corresponding growth in national funding of scientific projects up until the early 1970's. As Price remarks, the exponential growth of the population of scientists guarantees that there are more scientists living at any point in time than there ever have been in the times preceding them, in aggregate, with a doubling rate of 15 years, while expenditures on scientific research doubled every 5 years [52, 18]. Price also, incidentally, notes the development of understanding about the organization and mechanisms for the conduct of science, calling for the development of a formal study of science policy in the pursuit of theory, "that will do for science what economics does for the economic life of nations." Lew Kowarski describes the development of an understanding of Big Science, largely on the part of its administrator-scientist practitioners: Weinberg was a researcher and then director at Oak Ridge National Laboratory, Adams, who

penned “Megaloscience”, a director of the UK Atomic Energy Authority’s Culham Laboratory and later Director General of CERN, and Kowarski himself was a leader of the Data Division at CERN and technical director of the French Atomic Energy Commission. The end of this exponential growth, particularly of the growth of funding, is discussed with some concern, particularly about “disastrous oscillations” as a result of reductions in funding [93, 18]. Similarly worrying to these authors is the understanding that while the overall population of scientists is on the increase, that the number of great scientists that produce truly revolutionary work seems to remain the same [163].

Capshew and Rader [39] collect a broad swath of literature focusing on Big Science activities, noting that research in Big Science has traditionally ranged across a number of different fields and perspectives. The authors generally find that there appear to two main streams of big science research, the first on focused on the shift to big science and the consequences involved in growing ties between scientists and corporations, government, and the military, the second focused on the evolution of science into big science and its further development as “big” becomes “bigger”.

Within these two areas of thought, the authors identify a nine different characterizations other authors have developed to describe big science, each of these having more or less bearing on this work in some way: *(1) As Pathological:* Weinberg, the originator of the phrase, as well as Merle

Tuve [156], Norbert Weiner[164], and others, view Big Science as pathological, voicing serious doubts about the effects of the shift towards Big Science on scientific inquiry. The influence of money, purchasing, and administration is held to have an overall deleterious effect on science, both in terms of what is investigated and upon the investigators themselves. Weiner writes that multi- million-dollar projects are more easily funded than smaller investigations, to the detriment of "creative science". Similarly, Tuve states that the organizational model of Big Science reduces the capability for disciplined thought on the part of the scientist. Weinberg in particular notes the dangers of engaging in space exploration or high-energy physics while leaving pertinent social problems unaddressed. *(2) As Scientific Phenomena:* Price, with his description of the growth and eventual saturation of scientific growth, is engaged in understanding Big Science, and all science, as its own phenomena to be studied in its own right with the tools and questions that can be addressed at any other field. *(3) As an Instrument:* Capshew and Rader note the monumental scale of science as well as its interdependence upon technology, weaving together statements by both Weinberg and Price, noting that the use of the computer to automate and routinize activities previously handled by scientists results in data reduction and managerial skills coming to the forefront of organizing scientific activity, citing Galison [70]. By its own right, scientific inquiry is also largely an instrument of the state to achieve its ends, which are not bound to purely scientific goals. *(4) As Industrial Produc-*

tion: A number of authors identify Big Science with the mass production and commodification of knowledge products [171, 74]. This commodification means that science is also a market-conscious activity. Echoing back to Weinberg's concerns about the growth of administration, Ravetz [128] finds that engagement in basic research is essentially a large-scale capital investment, requiring the type of management activities of any project of such scope, resulting in the "concentration of power in the hands of science administrators". Furthermore, the process of doing science is also formalized and regimented, jarring with the notion that research is a craft activity, and removing boundaries that allowed work styles to vary widely. Industrial science also brings organizational impacts, requiring either scientists themselves or science administrators to make decisions about the formation of teams and disposition of morale or management issues.

The further areas of big science literature described by Capshew and Rader characterize extended problematics of Big Science: *(5) As an ethical problem*, Big Science represents a break between the responsibility of individual researchers and the ethical implications of their research. With a large number of researchers on any project, each contributing a small amount of work to a greater whole, accountability for the impact of research is diluted. A further ethical issue is created when scientists are given the authority to decide which projects are funded and supported, calling in to question the goal of research: is it to reveal knowledge, serve the state, better humanity? Finally, the scale of Big Science and the re-

sources required creates an obligation between the scientists involved and their funding organization, corporation, or patron. *(6) As politics:* the decision to engage in a particular research project becomes a policy decision, as increasingly any project involves organizational resources, spending, and tracking of research outputs. This means that Big Science projects, which make up a greater portion of all scientific activities, are in competition with each other for resources from science policymakers. Frequently peer reviews of proposals are conducted by fellow scientists, but the funding, management, and direction for scientific activities increasingly comes from corporate or military sources, with inherent policy questions. *(7) Big Science becomes institutionalized.* By their nature, these projects are large-scale activites with attendant expenditures and long-lived infrastructural elements. Not only are there physical and financial manifestations, but the operating organizations—labs, university centers, institutes, and the like—engage in institutionalizing activities, which include participating in coalition-building and inter-organizational cooperation, extending tasks and functions of organizations. Institutionalization of science projects means that the number of stakeholders increases, and that research becomes beholden to multiple dimensions. In my own research about cyberinfrastructure, I note that cyberinfrastructure becomes an institutional partner of researchers, with its own stake in the course of research [89]. *(8) As cultured activities*, Big Science projects investigate and put importance on the type of research that their participants

as well as their constitutuents find important, and make conclusions that are similarly informed. Finally, Capshew and Rader note studies such as Pickering's [123] which characterize Big Science as *(9) a form of life*. Scientific activities can have have very similar antecedents but be enacted with vastly different activities and techniques, indicating that there is much more at work than a programmatic process of scientific advancement.

Galison and Hevley [70] collect a number of articles which trace the growth of large-scale, collaborative scientific projects, as further and further inquiries into the nature of our world necessitate more intense scrutiny by teams of scientists from a number of disciplines. Some of this scholarly work focuses on the differences between Big and Small science activities and the implications that Big Science projects have for funding and scholarly output [70]. Other scholarly work on Big Science activities examines collaboration activities and the incentives for broader collaborative efforts [79], as well as the tendency of high-energy physics to be used as the exemplar of collaborative Big Science, when not all disciplines seem to follow Big Science principles. Some of the emphasis on arrangements within the high-energy physics community, focusing on the community of the culture and the informal collaborative collective [155, 91] appear to be less representative of science disciplines overall than they originally appeared. Rather than than a flexible, democratic, and flat configuration being the state of the art (and answering some of the doubts about Big Science projects turning scientists into factory workers), it appears that

interorganizational collaborations adopt a number of forms, which incorporate different levels of formality and aspects of classical bureaucracy based on specific needs of the project and interests and characteristics of the partners involved. An examination of the organizational structures of a number of physics projects identified a tendency towards traditional forms of bureaucracy when participants have a high degree of uncertainty about each others' actions; when collaborative work was largely similar across teams or when professional norms, collaborations adopt less formal organizations, such as that exemplified in the high-energy physics community [44].

Recent quantitative research into funding of projects both big and small finds that funding Big Science activities, as opposed to the modest investigations Weiner said were being pushed aside, has variable results. One study of grant-funded activities in Canada found the relationship between funding and impact in terms of scientific output is quite weak, and that higher funded researchers had fewer citations of their work than the best researcher of any randomly-selected pair, leading the authors to conjecture that "photo-opportunity science" was often less effective in terms of impact than many small projects which in aggregate could produce greater impact [63]. Another investigation of National Institutes of Health (NIH) funding at high levels found that review scores on larger projects (greater than $700,000 per year) failed to correlate with subsequent productivity in terms of publications [28].

As Big Science projects have proliferated, the accompanying need for computational resources to run simulations, analyze instrument data, collaborate with other scientists, and reproduce results has grown in a corresponding fashion[170]. As computer science projects develop in order to support other types of research, they seem to tend to follow the models of the projects which preceded them. Kowarski describes the incorporation of computers for physicist's work at CERN, noting that first computers were used to replace desktop mechanical calculators, then for managing and maintaining the proliferation of instruments which were rapidly increasing within the laboratory, then for simulation of results, and finally for cataloging and searching publications of physics texts. As data and processing requirements increased, Kowarski notes, it became more and more pressing for facilities to have links to high-performance systems that were frequently located at a distance from the laboratory[92]. The demand for computational capacity was such that universities were faced with a dilemma between funding computer systems or funding researcher travel to the accelerator cite or to another research institution with its own computer capabilities. With the growth of expensive, large-scale projects, it is impractical as well as detrimental to teaching new scientists to co-locate all researchers with the device producing experimental data.

The development of the use of computers at national laboratories in the United States is illustrative of the simultaneous development of big

science and cyberinfrastructure. Yood describes the development of the Applied Mathematics Division (AMD) at Argonne National Labs, and the rise of the computational scientist. As physical sciences at the lab became more and more reliant upon calculations and simulations performed by the computers in the AMD, the scientists at Argonne who became more involved with numerical analysis and developed expertise in translating from scientific theories into algorithms became the first computational scientists. These computational scientists found themselves in competition with the early computer scientists, who needed to use the same computer resources in order to develop their own theories and methodologies. As Yood details, during the development of computational capacity at Argonne, computational scientists benefited from a connection with the guiding principles of their own scientific disciplines, while computer scientists struggled to define their own field, even to each other. The physical scientists at Argonne's needs for computational capacity continued to grow throughout the period Yood examines (1946-1992), while the general community for high energy physics acknowledges a significant reduction in funding outlays in the early 1970's. This constant growth of computational needs has continued, as the size and density of data has increased, and the number of fields making use of computers in order to develop analyses has increased as well.

In "Image and Logic", Peter Galison [69] describes the incorporation of computers and the advent of simulations as part of a strategy to improve

upon the speed of processing instrument films, and outlines the differences between the Alvarez group and CERN physicists, Lew Kowarsky primary among them. As the quantity of bubble-chamber film rapidly outstripped the ability of individual physicists working in the lab to read it, Galison describes the brigades of young women scanners who were needed to identify particle tracks and the two groups' efforts to develop systems to assist or replace them. Noting that the amount of film would soon require more human labor than feasible or even possible, these two groups pursued different strategies of identifying and measuring particle tracks. The Alvarez group steadfastly retained focus on human intervention with machine assistance, to the point of creating scripts run by the machine which elicited human response. The CERN group, with engineers for reading machines and bubble-chamber physicists in two disparate groups within the organization, focused on the the development of pattern-recognizing machines, which would then provide results to the physicists. Galison notes the Alvarez group's techniques to compare machine-recognized findings with the pattern-recognition skills of physicists, and follows the disscussion of simulated experiments from early exchanges at CERN about statistical processing of data. In his chapter on the development of simulation techniques, Galison explores the "quasi-material dimension of material culture" embodied in simulations in physics, climate, number theory, and other disciplines. Simulations are quasi-material in that they not only drive the development of theory

but are the formative activities in building the instruments which inform the physical sciences – that is, as Galison notes, no instrument could be built without a simulation of its activities first being conducted: "Withough the comuter-based simulation, detectors like the TPC were deaf, blind, and dumb: they could not acquire data, process them, or produce results. Simulation, then, which Edwards describes extensively in his examination of climate research [54] is intimately involved with not only the computational end of research but also in preparations for building the instruments which record new data for analysis.

Hallam Stevens describes his time working at the Broad Institute and European Bioinformatics Institute and studying the practice of science and the rise of computational biology in "Life out of Sequence" [146]. As scientific computing grew in capability and utility in the latter half of the twentieth century, biologists for the most part refrained from use of computational techniques. As Steven writes, for a long time, the inquiries made by biologists simply did not fit well with the computational capabilities available. When gene sequencing techniques began to create large amounts of coded data, this data could be more tractably dealt with using computational means. Stevens contends that, in the process of creating a large stream of new data, and utilizing computer systems to analyze and compare sequences, computational biologists. These data streams were incorporated into first into text files, then into relational databases, and these relational databases were eventually federated into

multi-institutional banks. Stevens traces the development of biological questions along this course: when computers were able to examine flat text files, the questions and analyses revolved around individual genes; as they progressed to databases, the object of inquiry was about alignment; and with the rise of federated databases, multielement analyses began to be possible. Stevens conjectures that in the development of biological databases a form of theoretical biology arises, in which structuring elements are related to the basis of inquiry. As databases are created, their structures reflect scientific thought about the biology at work. Finally, Stevens notes that the immense amount of data involved and the techniques for engaging in analyses require the development of imaging techniques: for matching and comparison and for reducing data to more tractable forms–leading to the need for more advanced visualization techniques to provide useful ways of interacting with the data.

While for Weinberg, Big Science was characterized by monumental, aspirational projects, with attendant questions about the value of pursuing these projects, the notion of Big Science has been extended to any number of projects. These Big Science projects are often characterized by their collaborative nature, with institutions and sometimes disciplines working together to address a particular question. The pre-eminent Big Science project of the years around the turn of the century from the 20th to the 21st has unquestionably been the CERN Large Hadron Collider (LHC), which has engaged the work of physicists around the world at a

multitude of institutions. By the same token, the LHC has an associated cyberinfrastructure project to support the analysis and management of the vast amounts of data created by LHC instruments. The LHC Computing Grid has a similar scale in the realm of cyberinfrastructure projects as the LHC among instruments, an extremely large, distributed computational grid with requirements in 2008 of 140 million SPECint2000 units (a software benchmark produced by the Standard Performance Evaluation Company, or SPEC) 60 petabytes of online storage and 50 petabytes of archive storage. [13] Without a doubt, the LHC Grid represents an extremely large collaborative project in its own right, a multi-tiered system moving data across Europe and America in order to provide computational resources in a flexible fashion. As scientific projects have evolved, their computational needs continue to increase.

2.1.2 The development of collaborative cyberinfrastructure

Research on scientific inquiry runs across a broad continuum from examinations of individual inventors, to work with groups or labs, to the establishment of systems, and with the establishment of large scale science beginning around 1930, the continuum has been expanded to include these projects, which are characterized by questions from outside about their impact on public activities, about their expenditures, and their outcomes, but also questions from within about sharing credit, cooperative work, and how to incorporate differing scientific and cultural practices

into a large scale project [70]. Few would disagree that existing problems or "Grand Challenges" of science are all but inaccessible to the individual researcher, and that team-based projects, sometimes spanning hundreds of scientists across the world, are now the order of the day. A somewhat smaller body of research concerns itself with the material and administrative needs of these scientists, the laboratories and computers that make their inquiries possible, and the configurations and concerns that underlie the management of these large-scale projects: the supporting infrastructure. Star [145] notes in her discussion of studies of infrastructure that while the subject of inquiry initially appears dull, dramas of system-building lie underneath for one who would "study the unstudied".

Even as early as Lawrence's cyclotron labs , historians note friction between scientists and technical staff [70], which is carried on into newer sites, such as Argonne National Laboratory [170]. In Yood's account of the Applied Mathematics Division at Argonne, the computing center becomes a central location for multiple labs, where work is completed on computer systems in support of newly- prominent numerical analyses, but tensions between the center and the scientists abound. Edwards [54] discusses the work of numerical simulations of climate models, noting that the needs of the climate research community are focused around obtaining the fastest available computing systems, with demands frequently one step behind those of national laboratories. These computational, storage, and networking systems, the software that enables their use, and the people who

support them, make up cyberinfrastructure, also referred to variously as "grid computing" [64], "collaboratories" [61], or e-Science. While cyberinfrastructure and the researchers who make use of them have their differences, they are perhaps more alike than different, and cyberinfrastructure represents an extension of the infrastructural features that make up a large portion of our everyday lives, outside our notice until they fail.

While cyberinfrastructure is frequently presented in flashy trimmings and proclamations about novel and exciting approaches, it shares a number of features common to any infrastructural project. Infrastructure projects have considerable scale, knitting together separate, local systems into networks with a complex coordinating center. Infrastructures engage in transformative processes, not only of technology but of organizations, cultures, and rules, to manage spanning into new domains. Finally, the cooperation of local systems in the greater whole is facilitated by gateways, which make take the form of standards, protocols, or policies [55]. Cyberinfrastructure activities are characterized as enacting technology, organizing work, and institutionalization [129], a broad enough set of areas that many types of infrastructures might fit. However, there are som critical differences: cyberinfrastructure demonstrates considerable versatility over traditional infrastructures, they are reflexive in that producers and users frequently are part of the same global infrastructure, and in that organizations can examine their own components (software and data) as information within the infrastructure [125]. Star and Ruhleder [144] de-

scribe infrastructures for information architectures as having a number of dimensions. Infrastructures are embedded into other arrangements, social and technical, they are transparent, once in place we rely on them without thinking about how they are constituted, and by converse, they become visible only when breakdowns occur. They extend beyond a single event or place. Infrastructures must be learned by their community of practice in the process of establishing membership (which is a precursor to transparency), and they are influenced by and have influence upon their communities of practice, determining such things as disparate as working hours, user interface configurations, and more. Infrastructures, as means of binding together multiple locations or times are also the product of standards, without which none of the coordinating activity that constitutes them could take place. Infrastructures are built on pre-existing infrastructures, never implemented whole cloth, and they are never fixed in a global fashion, but only in modular increments [145, 144]. Some of the tendencies of cyberinfrastructure are shared with scientific collaborative projects as well, work based on the collaborative projects finds that teams working on cyberinfrastructure must have a means of achieving commonality across fields, there are issues with preparing younger members of the community to take over responsibilities defining boundaries of the organization and with each members home institution, and that frequently goals are emergent, rather than planned. Furthermore, project members must agree on how to represent their organization and its out-

puts to potential clients, funding agencies, and home disciplines [81].

Studying cyberinfrastructure, then, is one way of performing Bowker's [34] infrastructural inversion on the practice of science. While Paul Edwards examines the practice of climate science starting from the instruments and collectors of data who contribute to the models and analyses that make climate science works [54], the cyberinfrastructure investigator begins by examining the systems, networks, software, and people that support computational analyses in these large systems. Typical cyberinfrastructure organizations are large, composed of many hundreds or thousands of members, and do not easily lend themselves to observational or ethnographic methods, but may provide a wealth of documents and interactions that expose some meaning for the organization. Some ways of investigating infrastructures include distinguishing between a master narrative and others, finding work that is not immediately visible such as workarounds or adjustment activities, and examining both overt tasks and articulation tasks for users in order to find the explanation for unexpected obdurate barriers [145]. In terms of identifying a master narrative, cyberinfrastructure leadership often provides information of a particular tenor: managing cyberinfrastructure requires skillful arrangements of technology and other resources, with research providing the overall embedded activity that underlies cyberinfrastructure development, which may account for the research-like nature of many cyberinfrastructure projects. Furthermore, leadership notes the difficulties in communicating

value to stakeholders is a multi-level process, often involving the identification of second-order effects, and in obtaining and mentoring staff and leadership of centers [27, **?**]. Other research into cyberinfrastructure centers notes a concern with sustainability of projects, participants, and the incorporation of novel technologies into work of existing users [129]. Considerable work has been done around the negotiations between different groups that allow for work practices to be completed (articulation), including the importance of interpersonal coordination over formal organizational structures, the process of negotiating trust between multi-group projects, and dealing with uncertainties [60, 97] [25, 62].

Cyberinfrastructure is constituted of technological resources, networks that connect them, software that enables work to be completed on them, and people to architect, implement, and support them. While traditional infrastructure and cyberinfrastructure have a number of similarities, the flexibility of software, considerable flexibility of organizations, and the dynamic environmental concerns that shape it make cyberinfrastructure an extension of infrastructure. Like infrastructure in general, it is a complex system, with multiple parts, and large collections of stakeholders and participants. Like any other scientific instrument, cyberinfrastructure is created to facilitate the output of scientific products, and for computer scientists, cyberinfrastructure may be an instrument. It is not created out of nothing, but is built upon the infrastructures which preceded it, and configured based on demands from scientists who want to make use

of it and funding agencies who pursue particular standards and goals of implementation.

2.1.3 Literature on cyberinfrastructure usage and scientific production

A few resource utilization and input-output studies have been conducted specifically for the TeraGrid and XSEDE projects. Using the TeraGrid database of accounts, allocations and CPU charges, Hart [80] examined resource utilization on the basis of individual workload characteristics, finding that patterns of job submissions to TeraGrid systems do not necessarily correspond to usage modalities, that is, submitters such as gateways that might be expected to submit jobs across a wide range of resources, frequently submit jobs to a single resource rather than taking advantage of the grid as a whole. In contrast to a true grid, in which middleware submit jobs across a wide range of available resources, TeraGrid submissions are largely user-dependent, and they largely reflect the usage policies in place at a particular site. Hart also finds that allocations are successful in controlling the global demand across a number of systems. HPC usage across a federation of systems appears to reflect usage patterns previously displayed on single systems, but the manual balancing of the TeraGrid allocations system creates different patterns of usage on the individual systems. Another study by conducted early in my own research [88] documents the results of a network analysis of TeraGrid user

and project allocations. Results show that large projects frequently make use of multidisciplinary teams and that teams working on TeraGrid allocations are made up of both domain scientists and technical specialists, while smaller groups tend to be populated by domain scientists alone. Computer scientists may be members of a number of projects in varying scientific domains, while domain scientists tend to remain in their area. Later, Furlani et al, creators of the XDMoD project whose data I make use of in this project, used information on resource utilization to improve operations by analyzing resource-specific data on throughput together with codes on individual elements of XSEDE, to characterize project activities, and to identify under-performing codes [68]. The authors show that molecular biosciences are rapidly gaining prominence in the TeraGrid and XSEDE environments, and that they represent a significant departure in usage modality (many cycles on a smaller number of cores) as opposed to traditional HPC domains such as astronomy, physics, and atmospheric sciences, in which large analyses are employed that utilize a large number of cores.

A number of measures for improving impact assessment for the use of TeraGrid for its scientific users are proposed by Sheddon, et al. [141], which notes that the current measures (such as number of users, usage data, and publication information) provide information about outputs of the system, but not necessarily scientific outcomes. This team, established as a "Requirements Analysis Team" by TeraGrid leadership in

order to ascertain requirements that would extend and improve TeraGrid activities, recommended a number of activities that would capture impact on scientific research and knowledge, including improving the proposal system in order to better capture data such as supporting grant funding, adopting the NSF's practice of keeping a database of "science nuggets" (short description of scientific work done and the contribution of the TeraGrid to the project), and improving survey practices.

Moving to the publications of the XSEDE user base, Wang et al follow a bibliometric analysis of publications supported by TeraGrid and XSEDE allocations, describing the impact of resource utilization on publication frequency [160]. Results show that while at the individual project level, the use of TeraGrid and XSEDE infrastructure does not show a strong positive correlation with impact metrics (e.g., number of publications, number of citations, h-index, and g-index), when usage is aggregated at the field of science (FOS) level, larger allocations are positively correlated with all four of these impact metrics, leading to the conclusion that resources matter in terms of consumption at the aggregate FOS level.

In efforts to understand linkages between regional scientists publication and citation, Mazloumian et al categorize the inputs and outputs of research work based on citations and publications with a focus on the exchange of information across national boundaries [105]. The authors identify knowledge sources and sinks by geolocation, and find that the coastal United States, England, and parts of Central Europe appear to

be knowledge sources, while Asia and South America appeared to largely be knowledge sinks (researchers citing others in their publications but not being cited themselves). This geographic exchange of scientific knowledge shows that flows of information can be mapped in order to identify sources and destinations of scientific information.

2.2 Science Policy

Funding in the form of grants is commonly regarded as the engine of scientific progress, providing the materials and labor required to enact the ideas of the nation's thinkers and researchers. Grants for research involve serious investments on the part of the United States government, constituting $7.3 billion for the National Science Foundation [113] and over $37 billion for the National Institutes of Health [112] for overall budget requests, of which the greatest bulk is dispersed for grants for scientific research. Grants for research represent an unusual facet of government partnerships with other organizations, posing particular questions about monitoring, outcomes, and the process by which research grants are awarded and kept. Analysis and measurement of the processes of research is not an easy task, compared to other services paid for by government, and yet the need for support of scientific research has been emphasized since the end of the Second World War [36] . Further clouding the picture, a number of current "Grand Challenge" projects [42] have been identified for specific targeted research investment that will require sub-

stantial collaborative efforts across multiple institutions and disciplines, creating an environment of where cooperation and competition mix together. Given the need for educational and economic improvement that is driven by the products of scientific research, and the recent political climate in the US and Europe, which seeks to cut funding in the interest of generating savings, understanding the processes by which grant funding for science works and how the best results of science policy may be obtained is critical to justifying the the continued investment in scientific research, and in improving return on those investments, which yield impacts that affect the well-being of nations.

In this section, I explicate the processes of grant funded research in the United States and examine characteristics of the current system of funding in comparison to the more common practice of contracting out services, describing some of the key differences in the research process that necessitate grants as opposed to contracts, and make clear the the use of grant funding for scientific research as opposed to other methods of funding. Secondly, I examine the most common form of selection and monitoring in the grant funding process: peer review of grant proposals and science products. I describe some of the benefits and shortcomings of peer review and the relationship of peer review as a performance management tool in the context of new initiatives for monitoring performance in recent years. Finally, I provide an overview of the system of competitive processes for funding cooperative research, looking for the effects of this

system on the processes of research.

2.2.1 Government funding for basic research

The United States Government has long used grants as a particular mode of completing work in order to support and stimulate particular activities that it wishes to accomplish. Grants are distinguished from direct activities of government in that they are executed by outside entities, while also different from indirect activities such as contracts, loans, and regulations in that grants tend to be less prescriptive about the process of the activity and the monitoring that can take place during grant execution [132]. Scientific research is particularly suited to grant funding, as it tends to be an open-ended and in many cases, the final applications may be very far from the initial exploratory research. In contrast to contracted projects or even applied engineering projects, scientific research is intended to answer particular questions or provide specific information about some previously unknown quantity. A contracted project generally provides a particular service or good that is usually well-understood and has identifiable, concrete qualities that define its performance in some way or another. Likewise, an engineering processes leverage existing knowledge in order to create a previously nonexistent thing (a bridge, a computer, a rocket engine) with some defined characteristics, tolerances, and performance that the finished project must meet. Governments fund scientific research to get answers about some operation, phenomena, or interaction

in order to provide information on which to act, either in the sense of informing policy decisions or in the sense of informing the engineering of technologies that take advantage of the knowledge obtained in research. In this sense, the stock and trade of research is in unexplored territory–making it difficult to provide timetables or define outcomes. The answers that research may provide may be significantly different than expected at the outset of the investigation. At the same time, science does have a generalized productive function in that most researchers are expected to provide at least a timetable by which they can expect to have determined an answer, or determined that further information is necessary in order to arrive at an answer.

Historically, funding for research can be traced back to the pioneering efforts that took place in the French Academy of Sciences from the late 18th Century up to the First World War. Work on the history of grant-funded activity traces the shift from prizes rewarding researchers for past work to the formation of the academy's funds for supporting proposed research [49] . Hanson [78] details the contrast between prizes and grants in his examination of McClellan's [106] history of scientific societies. Hanson cites principal-agent theory in explaining the shift from paying for results in contrast to paying for effort. As principal-agent theory states, paying for results is most successful when the principal can easily specify the results and is risk-averse, while the agent is able to take on more risk and raise capital. Paying for effort is attractive to the principal when effort is

easily monitored, when there is more information to guide the selection of agents, and the principal is prepared to take on more risk. In this way, Hanson describes the shift between retrospective and prospective forms of funding for scientific research, as British and French governing bodies began to reduce the amount of rewards funding past research in favor of grants that would sponsor later research. The modern era of research funding began when Vannevar Bush embarked on his program to extend the funding and efforts of wartime research bodies into the National Science Foundation [100]. At this point in time, university research programs are reliant on federal money in order to carry out their agendas [127], and federal funds support more than 60% of academic research [**?**]. In the overall picture of research funding, about 200 higher education institutions spend roughly 95% of US federal funding for research on campuses. Other areas of government investment include government laboratories, which receive roughly one-third of federal expenditures for research and development, and federal contractors, which partner both with government laboratories and research groups at universities and colleges [47].

In describing the modern grant funding environment, Beam and Conlan in [132] detail a number of features of grants that make them particularly suited to funding scientific research. Grants are gifts with the intention of supporting and fostering a particular activity. The grantor is able thus to participate in service provision, while leaving the details of service provision to the grantee. Grants do not need to be paid back, and

do not require a particular product to be delivered to the donor, but they do represent a structured relationship in that some kind of activity (as in Hanson's paying-for-effort) is stimulated with an end to creating a situation that the grantor wishes to arrive at at some future date. Selection of grantees is based upon a number of characteristics often depending on the domain of the desired activity. In the scientific realm, peer review is the principal means for selecting grantees to be funded, retaining a tradition of being the gold standard for scientific evaluation, despite limited transparency of process [45] and has been accused of impeding innovation [17]. Beam and Conlan in [132] note that grants rely on the administrative apparatus of the grantee to manage implementation and execution of the process, so that the grantee has a much greater degree of discretion in comparison to contracted services. By the same token, grants tend to be highly visible areas of action. Large grants are typically covered in press releases and an individual scientist's track record in winning grants is a determinant of later grant awards [17]. This visibility also serves to make grants for scientific research a target for cost-cutting measures by Congress [159]. Since the end of World War II, grant funding has been awarded to fewer individual researchers and more centers and collaborative projects, as many have noted [127, 72].

Grants share a number of features in common with government-funded contracts. Both forms of funding can have a financial commitment on the part of the contractor or grantee. For contractors, this may take the form

of some kind of initial investment or a bond ensuring completion of the contracted work. For grantees, this takes the form of so-called "matching funds", which are dedicated funds contributed by the grantee to accomplish some portion of the work. Both contracts and grants feature recurring renewals of funding in order to continue the sponsored activity. This allows for the activity the government wishes to foster to be extended, with the partner who possesses the greatest experience in carrying out the activity. Both grants and contracts are frequently awarded based on the recipient's previous track record and expertise. Liebert [99] examining the determinants of receiving group funding, finds that previous productivity is the largest antecedent for receiving grants. Likewise, Leroux [98] states that experience and track record is one of the primary elements of bid evaluation in the contracting process. Finally, dependence on government funding in the case of both contractors and grantees leads to a sensitivity to political factors that would otherwise be absent [72], as organizations are affected by changes in agencies, administrations, and politics that affect their sponsors.

The differences between grants and the contracts that have become so popular for completing government work in the last two decades fall in the area of goals, monitoring, and specification. While one of the most important facets of contract work is creating the specification of work for the contractor to complete [98], grant solicitations for research work are frequently limited in specification, by necessity to topics of interest or to

specific research questions. In the world of research it is often unclear which tactics can be used to arrive at an answer, and novel solutions tend to be the rule for getting to new knowledge. As such the scope of work, budget, and changes to the research program tend to be more flexible than in government contracts [157]. Frequently research requires exploration in order to understand the domain under study before embarking upon a research agenda. Monitoring is also different for research grants. In grant-funded situations, monitoring conducted by the grantor often consists of gathering information about the activities that the grantee engaged in during the period of the grant, rather than delivery of results [132]. Vannevar Bush in his report supporting the creation of a national organization for supporting scientific research noted that the returns on investments in basic science would greatly and certainly outweigh any failures along the way: "Statistically it is certain that important and highly useful discoveries will result from some fraction of the undertakings in basic science; but the results of any one particular investigation cannot be predicted with accuracy" [36]. Grants also tend to be paid in annual installments, rather than based on reaching milestones or delivering particular products, likewise reporting for grants happens on an annual or periodic basis rather than the frequent reporting schedule that is common for contracted work, and finally discretion for leadership in the project is largely in the hands of the grant's principal investigator rather than the government supervisor [157].

Performance management for grants is perhaps the most important difference between grant funded work and contracting work, and monitoring of scientific progress has singular difficulties. Partha and David [121] catalog the difficulties of economic evaluation of research: economic returns may come quickly or may take decades to realize, rights to intellectual properties are difficult to extract economic rents from (in fact restricting access to research may hamper further returns on initial investments), fundamental research progress may have dramatic and far-reaching impacts that are difficult to capture, and it is especially hard to forecast the success of any one particular research project. As noted above, Hanson's [78] appraisal of focusing on effort as opposed to results has marked the transition from measuring research outputs towards measuring research processes. The standard operating procedure for evaluating both inputs (proposals) and outputs (scientific work) of grant-funded projects remains the peer review process. Garcia and Sanz-Menendez [71] discuss the context of peer review as metric of scientific research quite fully in their evaluation of competition in research initiatives. The authors begin tracing the path of peer review with the assertion that individual reputation and credit are central to the creation of the social structure of science, and that recognition by ones peers is the foundation of legitimacy and leadership in a given field. Garcia and Sanz-Menendez note that the measurement of scientific production has long been based in volume and quality of scientific publications, but that these metrics cannot be sep-

arated from peer review. Peer review, despite some of its flaws outlined below, is not only the basic mechanism for ensuring quality of research, but also a critical factor in monitoring the efficiency of government investment in science. Peer review provides legitimacy to governmental bodies, and scientific work which has passed peer review has greater esteem in its scientific surroundings [71]. However, with the advent of new initiatives in government for assessing and monitoring of performance, peer review has had mixed fortunes as an evaluative tool for officials in charge of awarding research grants.

2.2.2 Making Grants Measurable

Shapira and Kulhman [140] describe the growth in requirements evaluation of research projects as governments attempt to control costs and derive better benefits from programs, noting that there are significant issues to measuring performance in this area. Impacts of these programs tend to be diffuse, as do costs, leading to difficulties in capturing all of the inputs and outputs. As research programs grow in complexity, including more disciplines and addressing broader problems, the evaluation of these programs must similarly become more complex. Government demands for continuous monitoring and program learning initiatives for research have led toward inclusion of subsidiarity, socioeconomic effects, and broader impacts into research evaluations. The increasing frequency of public-private partnerships for research also increases the complexity

of program evaluation [140]. Partha and David [121] characterize the the new attitudes towards measurement of research projects as a new economics of science, in which the previous free-market scientific workplace, characterized by scientists competing in the peer review process in order to gain funds and recognition is supplanted by a more interventionist government hand that is in the process of turning science toward more applied tasks. Government demands for better program evaluation both in the US and in Europe, as well as budgetary constraints from the recent economic crisis, have resulted in a call for more scrupulous examination of research performance.

The response to this call for increased evaluation and measurement of performance, has been variable at best. Cozzens in [140], providing the context of evaluation in US research funding, describes the clash between the traditional evaluation tools for research, peer review and the journal selection process, and the new requirements based on the Government Performance and Results Act (GPRA) and increased requirements for management performance from the OMB. Peer review as the status quo for evaluation of science works in what Cozzens describes as an "autonomy-for-prosperity" model. Agencies support research activities in order to solve specific problems in an indeterminate amount of time, with limited oversight from Congress or agencies. Emphasis in evaluation is placed on the input end of the process, based on the quality and relevance of research proposals, and most importantly the accountability

of this evaluation is placed on the research community, who is responsible for making decisions fairly, rather than on the researchers themselves to produce results to the general public. Guston [76] notes that peer review makes up a substantial amount of the selection process for research: $37.7 billion or 86% of the reported total funding for research is merit reviewed. Applied research agencies, in contrast, have review processes based in personnel evaluation and budgeting that determine quality, although Cozzens, Bozeman, and Brown [47] note that there is a shift towards the competitive model of peer review even for these agencies. Peer reviewed grants are a feature of new federal research funding programs in the Department of Agriculture, the Environmental Protection Agency, and the Advanced Technology Program [76]. Peer review for research projects can happen both prospectively in the proposal selection process as well as retrospectively in evaluation [116], and have also been used as inputs in evaluating information for drafting regulation, creating state policies, and in evaluating courtroom decisions [76].

The peer review process conflicts directly with GPRA requirements for monitoring outputs of research, which explicitly focus on planning and achieving strategic objectives, rather than a culture of fairness in evaluation of proposals. As Cozzens in [140] states, this "clashes with the traditional notion that the benefits that flow from research cannot be predicted in timing or content, but rather are visible only retrospectively". Response to the new evaluation requirements has been met by providing

measures that are generic and qualitative: outcomes such as “advances in knowledge” or “ideas, peoples, and tools”, or the NSF’s frequently sought-after “science nugget”, used to provide Congress with information about program success in a brief, easily-digestible package. As a result, such weak measures of evaluation lead to reinforcement of existing political forces, especially when coupled with another popular new metric of stakeholder input, which gives greater voice to those parties already engaged in the selection process [140]. Another approach to evaluation is to provide broad indicators of research progress: publications, funded research, and patents. Campbell in [38] directly contrasts the peer review and indicator approaches finding that peer review results in complex but subjective evaluations of research work, while indicators are objective and easily quantified, but tend to be superficial in nature. Hagstrom [77] notes that peer reviewers frequently are able to identify the authors they are reviewing, or at least make educated guesses based on prior research and citation patterns. There is some evidence that researchers understand the peer review process and anticipate elements of it when drafting proposals. Knorr-Cetina [90] found in comparing proposals submitted for peer review to those without peer review that the style of the proposals changed rather than the content of the science inside. Furthermore, peer review is frequently conducted by established researchers, which leads to a problem in the assessment of new and innovative research directions, and relationships of mutual dependency that create self-reinforcing factions

within scientific communities [38]. The world of the peer-reviewed scientist may be viewed as one mired in competition with other researchers – first to get research proposals approved in order to get funding, and then to get the results of that research published.

2.2.3 The competitive element

Competition is in many ways the coordinating feature of scientific progress just as it is in economic activities. Hagstrom [77] describes the competition that takes place between scientists as specifically occurring when a scientist finds that her research in a particular area, on a problem not previously published, has been beaten to publication by another researcher. This form of competition may be extended to include being passed over in favor of another researcher in the grant selection process. Latour and Woolgar [94] established research funding as a vital part of researchers' credibility and reputation with other scientists. Garcia and Sanz-Menendez [71] sum the idea of competition up well: "Thus, competition for funds is an essential mechanism in the cognitive functioning of research, articulated in the credibility cycle, and a vehicle for relationships between science and government". Competition between researchers has a number of valuable features that aid scientific development. Competitive publication practices mean that additional researchers may be working on the same problems, which ensures an abundant supply of possible investigatory techniques and results. Competition drives hard work on the part

of the competitors to outdo each other. Finally, competition reduces the risk of dilatory publication, and it encourages differentiation and innovation as scientists attempt to identify new problems to explore [77]. While competition should promote the best quality research, issues have been identified with competitive processes for publication and funding that may slow the progress of science. Competition thus has a complex relationship with peer review. Laudel [95] notes that the competitive process may have impacts to the course of science as scientist averse to risk select other research topics in order to avoid competition and increase favor in peer review, promote mainstream or existing research techniques in order to be more competitive with particular review boards. Reports from leaders in grant-funded research centers find that competitive resubmissions for funding has a disruptive effect on getting the work of the center done [143].

In the case of Supercomputing Centers and the TeraGrid and XSEDE projects, funding cycles create periods of collaboration and competition. One respondent during the TeraGrid mentions "coopertition": tighter cycles between NSF solicitations being released and the performance periods that centers are engaged in create situations where staff are expected to work together to make one project a success while preparing competitive proposals against their coworkers for the next cycle of awards. Other cyberinfrastructure experts in the same study state that cyberinfrastructure is not built in a three-year investment, but longer time-frames are

required in order to create robust, high-quality infrastructure that can support activities over the long term [172]. None of the current supercomputing centers in the US are run as facilities programs by the NSF, although they survive by engaging in multiple grants for cyberinfrastructure activities, which are receiving less funding for their activities. A pair of center staff worked out that recent solicitation during the time of the XSEDE project stated specifications for a system with a total allowable budget that would not support the electricity needed to run the system over the life of the grant, if it were awarded at their center. One of the respondents to this study jokingly refers to the situation, noting "the amount of money you lose on each individual award is made up for by the lack of frequency in awards coming out". Clearly the center heads feel that they are in the situation of subsidizing the activities of NSF, and most agree that the need to diversify funding sources by participating in multiple grants for research activities, while simultaneously competing for infrastructure projects which allow them to have the resources which give them credibility and weight in the cyberinfrastructure community [172].

2.2.4 Public Management Networks for Service Delivery

Alternative delivery structures in the form of networks, cooperative arrangements, and institutions have become the de facto standard for service delivery for the US government (as well as many western nations), for a number of reasons. Understanding these arrangements helps us as

scholars determine their effectiveness, the types and varieties of arrangements that may exist, as well as the ramifications for administrative arrangements that cross organizational or sector lines, to public representativeness, to responsiveness to constituting bodies, and to other networked organizations. In order to understand the XSEDE's environment, it is necessary to have a firm grasp on the concerns of the stakeholders who are investing in the XSEDE project and other computational infrastructures. This section explores some of the public management theory about networks for provision of services. While a limited amount of research here relates to information technologies, it does describe the conditions and constraints that determine the activities of the NSF, the NIH, and other agencies which support both basic and applied research in the US.

Cooperative arrangements are characterized by member organizations that are oriented towards a particular purpose (although not necessarily on all points), that work in concert, but without hierarchical authority over one another, without formalized relationships or group practices, and usually with limited coordinative power. These structures can be from the same sector or from a mix of sectors, and they can be aligned by common purpose in one space (to solve a particular problem) or in individual spaces (to solve the same problems locally).

The first component of evidence supporting Frederickson and Smith's [67, 66] claim is the funding provided by the federal government. As little as 15% of federal expenditures are direct funding of government agencies,

which is indicative of grant and contracting arrangements on a massive scale of operations. Some of these arrangements for particular goods and services are governed by formal contracts, to be sure, but as Scott and Davis [138] describe modern organizations, the bulk of these are also de facto network arrangements between multiple specialized producers. O'Toole [120] provides important insights about why networks (from this point on, I will refer to these alternative delivery structures as simply "networks") are increasingly prevalent. O'Toole notes a number of features of networks that make them appealing for service delivery arrangements. First and foremost is the tension between the simultaneous public demand for less government and increased services. This conflicting demand requires governments to do more without creating additional bureaucratic agencies, so arrangements with non-governmental bodies must be created. If citizens distrust bureaucratic structures, the more palatable policy solution may be the creation of a networked structure. Secondly is the requirement of government to solve wicked problems. Wicked problems are complex, multi-dimensional issues that cannot be solved by a single agency due to the interrelation of factors that create them [131]. Some of the wicked problems that governments wrestle with are poverty, illegal drugs, and now on a scale never encountered before the first decade of the twentieth century, terrorism. These issues have multiple causes, moderators, and suggested solutions, and no single agency is capable of handling all of the dimensions, therefore networks of agencies, private

firms, and non-profit groups with some kind of aligning characteristics work together. Networks also arise in order to deal with resource constraints, funding, or jurisdiction issues, if multiple localities do not have sufficient funding or power to create an over-arching organization to handle all of them, they may create individual organizations that operate in a network format in order to address issues. Finally, networks come together in order to coordinate the work of multiple specialized individual organizations to make the best use of expertise, resources, and experience.

Agranoff and McGuire [20] maintain that networks are not supplanting bureaucratic organizations, but rather that networks are made up of existing bureaucratic organizations that become members in networks. Generally in this case, responsibilities and activities are not devolved to the networks at large, they remain with the bureaucracy, with each member focusing on their own responsibilities and internal management issues. Klijn and Skelcher [87] examine some of the democratic implications of network structures, noting that networks tend to be managerial in focus, rather than representational. Networks exist in order to accomplish a particular purpose, rather than to create representative policy initiatives per se. Individual policies govern the activities of each network member, rather than an overarching policy that covers all members. In addition, despite their fluid nature (in comparison to hierarchy), networks tend to be largely stable, with few members entering and exiting

the group. That being said, the resilience of a network may depend on the "tightness" of bonds between members and what maintenance activities go into the network. One study of loosely networked managers found that apart from staff who acted as network intermediaries public administrators spent less than 20% of their time on working with the network and collaborating with network partners [20].

In terms of understanding networks, public management is largely starting with basic scientific principles and and attempting to identify concepts similar to management principles already examined by the field. Scholars [20, 119] call for an identification of the variations of network types, their scale and span. It would be useful to understand what kind of life cycle networks exhibit, how they form, how they are maintained, strengthened, or weakened, how they fail. Public management researchers are interested in the activities of network management that are similar to activities in hierarchical settings. Agranoff and McGuire have proposed analogues to hierarchical management such as activation and reinforcement, which are motivating / maintaining roles within network management. The study of public management networks also needs to explore the unit and level of analysis in order to be able to make claims about generalizability within or across network settings. It is important to understand how decision-making takes place in the network context, what effects nearness and farness of network neighbors has on choices for public managers. Scholars of public management need to understand

what the mechanisms of coordination are in networks, what manifestations of authority may be present in networks, and how to measure performance of networks. One possible source of authority in the absence of rational-legal authority may be reputational, which has its analogues in other examinations of socially-networked organizations. Researchers of public management networks should also be prepared to explore what role technology plays in network coordination.

Milward and Provan [108] provide comprehensive guidance on empirical study of public management networks in their examination of four mental health networks. In comparing the four networks, they look at levels of cooperation and complexity, the levels of activity, and the outputs of the networks organization. Provan and Milward identify antecedents for outcomes as especially important – what you have to work with determines what you are able to get done. This study provides concrete guidance on comparative network analysis, by providing variation in the observed organizations, comparing them on multiple dimensions, and attempting to identify output conditions.

Berry et al [31] explore multiple streams of research in relation to public management networks in order to identify some useful directions for research. Social network analysis, as pioneered by [110] and formalized with the use of graph theory [161] provides some potentially useful empirical tools with which to analyze and understand network arrangements. Social network analysis focuses on ties between actors and mathemat-

ical representations of these ties, the number of connections particular network nodes have, their directedness, and properties of the graph of all actors and relations, such as the maximum geodesic distance across the graph, graph density and centrality, the size and properties of the largest fully connected subgraph. Two of the useful techniques to come out of social network analysis in recent years are positional analysis and latent cluster analysis. Positional analysis relates to the relationships between small groups within networks (dyads and triads), and the likelihood of similar relationships within these small groups within the larger network. Latent Cluster analysis identifies groupings of actors within networks based on the length of ties to each other and the similarity of connections each actor has. It is important to note that one of the tenets of social network analysis is that analysis pertains to individual networks, and can tell researchers about features within the networks, but social networks analyses are not comparable across networks. Networks within the political arena may serve to give public management network theorists an understanding of the effects of influence within networks, and identification between network members.

O'Toole and Meier [119], theorizing about the "dark side" of public management networks, present a number of issues with network management that affect public management networking specifically. These issues include cooptation, shielding or masking of power, and deferment of unpopular decisions. With the relegation of policy responsibilities to

networks, government places an obscuring organizational level between it and the public. Where hierarchical channels are not existent and formalization is minimal, the connections between network members are not clear to legislative oversight, or visible to the public. Meier and O'Toole propose that this allows for members of networks to exert undue influence on the rest of the network in the form of cooptation. Cooptation of a network changes the network focus from its original area to one that is more preferred by the coopting member. The member can then coerce or elicit activities from other members that further its goals at the expense of the group goals, or network members can deliberately create confusion within a network and reduce focus on the goal so that less activity is carried out. One example of this may be the cooperative activities between the mining industry and the department of mining that led to lax safety procedures exposed in investigation of recent mining accidents. Members can also make use of this obscuring of official connections to exert power, or make use of the network identity to assume authority and accomplish initiatives the individual member would not be able to accomplish otherwise. Networks also insulate government in two ways. The ambiguity of network arrangements insulates the decision-making process from the public realm, and they insulate the responsibility of the government during implementation. A government agency may pass responsibility for decisions and actions on to the network as a whole, thus avoiding the scrutiny of oversight and public officials. The removal of public oversight

from the decision-making process is exceptionally worrisome to the authors, as it constitutes a reduction in democratic participation within the policy process.

In order to demonstrate the efficacy of networks, public management network research needs to identify and test analogs to structures in hierarchical public management. Researchers must create taxonomies of varieties and sizes networks, and perhaps even borrow from ecological models of organization to understand the population of networks. Further research on resource utilization in the network context is required. Further questions on networks revolve around practices of coordinating and promulgating network activities: What are the management activities that take place in networks? What are the channels and structures of influence and coordination in networks? What are conditions of successful performance? Moreover, if networks are structures designed to deal with wicked problems that hierarchical bureaucracies cannot, what defines success, which has yet to be seen with traditional forms of management? In order to understand the network life cycle, antecedents of network operations, and network performance, researchers need to examine networks longitudinally. Researchers may make use of social network analysis techniques to perform relatively complex analyses of networks, even with limited members.

The relevance of public management network theory to XSEDE is in the effective execution of agency goals through collaborative networks of

supercomputing centers. Some specific work on return on investment to NSF has already been done in order to compare between the overarching structure of XSEDE and providing the same services through individual (and competing) centers [150], finding that compared to individual centers, the NSF receives significant value from a coordinating organization. Some further exploration into similar programs such as NSF's ACI-REF program or the iPlant collective is warranted. Investigation into these activities can also benefit the public management sphere, as opposed to community or regional networks for human services (where many studies of networks have been focused), these projects are highly specialized and make considerable use of information and communication technologies in order to increase span of coordination and efficiency, and are exploring unique ways to use resources efficiently and effectively [43]. XSEDE and projects like it also are embarking on a number of initiatives in order to better ascertain performance and provide improved performance measurements to the NSF and to the legislature.

Chapter 3

NSF-funded cyberinfrastructure prior to XSEDE

The NSF's initiatives to support computing have been a long-term evolution of infrastructure. In this section I detail the development of the Supercomputing Centers Program, which marked a transition from computer systems as research project in and of themselves to infrastructure intended to support computational science research. Following that is a description of the initiation of the TeraGrid project, which replaced the Centers Program as the main provider of computational resources for basic science from the NSF. I trace some of the challenges and innovations of the TeraGrid, including researchers' examination of the tensions and dynamics of the TeraGrid project, which, owing to its structure and funding patterns, was a contentious project. Nevertheless, the TeraGrid managed to arrive at solutions for providing a distributed grid infrastructure for science. I contend that, thanks to the network organization of the TeraGrid and the development of other related organizations during the three major programs that funded and extended the TeraGrid, a vibrant cyberinfrastructure community arose. The professionals who had worked in

the computational sciences, information technology, and other fields, began to create their own network of those knowledgeable about research HPC projects.

3.1 The NSF Supercomputing Centers Program

The collection of resources and services currently provided by the XSEDE framework saw its genesis in the NSF Supercomputer Centers program started in 1985, and continued its evolution through the Partnership for Advanced Computational Infrastructure (PACI) and the TeraGrid programs. Each new program brought major advancements in the design and usage of cyberinfrastructure in support of research. The Centers program established a set of supercomputing centers which would lead the development of large systems [1]. From the mid-1980s through the 1990s, the PACI program in concert with the NSFNET initiative began the work of connecting these centers in order to provide access to computational resources to a larger group of researchers [6]. Starting in 2001, the TeraGrid developed a central organizational structure for obtaining accounts and allocations on systems, and worked to establish common user environments, including a standard software set, as well as providing services to help the optimization and utilization of scientific software, training, and education programs.

Much of the history of the NSF's support of researcher access to computing capability has been tied to the development of NSFNET and subse-

quently to the Internet, and accounts of the Centers programs are often closely tied to the development of the Internet and the Mosaic browser developed at University of Illinois Urbana-Champaign. Narratives about the development of the Internet and the World Wide Web frequently take center stage, but details about the development of the centers that these networking initiatives knit together are available.

Early support of computational systems by the NSF was focused on the development of computers themselves, rather than systems which would support computational research. As computer technology became less experimental and more of a component of research, access to computational resources began to become an issue for the development of computational sciences. The 1982 report of the Panel on Large Scale Computing in Science and Engineering, informally known as the Lax report, commissioned jointly by the DOE and NSF, noted specifically the difficulty of access to supercomputing resources to both research and defense organizations, as well as the development of more powerful computing resources as central to the research goals of the nation [96]. For the initiation of its supercomputing program in 1984, the NSF purchased access to supercomputer time at six sites: Purdue University, University of Minnesota, Boeing Computer Services, AT&T Bell Labs, Colorado State University, and Digital Productions [6]. This access program provided researchers with access to some of the most advanced computing facilities available at that time. In 1985 the NSF increased its support of computer

access by contributing to the support of the establishment of four supercomputing centers: the John von Neumann Center at Princeton, San Diego Supercomputer Center (SDSC), National Center for Supercomputing Applications (NCSA) at University of Illinois Urbana-Champaign, and the Cornell Theory Center (CTC). These were followed by the funding of an additional center at Carnegie-Mellon University (jointly administrated with Westinghouse and University of Pittsburgh), the Pittsburgh Supercomputing Center (PSC). These centers, together with the National Center for Atmospheric Research (NCAR), became the first endpoints on the NSFNET backbone.

In 1989, the NSF renewed funding of PSC, SDSC, NCSA, and the CTC. This arrangement held for 8 years until 1997 when the NSF restructured the program, retaining funding for only SDSC and NCSA. The NSFNET model is informative in that it provides the first major step of integration which required the separate centers to begin the process of standardization on networking protocols. This integration for networking activities presaged a larger push towards the provision of a common infrastructure with the mission of providing flexible enough capabilities to support a wide range of analyses while still providing advanced capabilities for resource-intensive analyses. In contrast to this integrative approach of the NSFNET model, the de-funding of CTC and PSC resulted from those centers' focus on continuing to deliver production science systems, rather than providing next-generation computer systems. These tensions between research

into the edge of computational capability and the provision of production capability echo the issues Yood found at Argonne, writ into a larger framework. As we will see, the NSF continues to balance the needs of computational sciences with development in large-scale HPC systems.

In 1997 the NSF restructured the Centers into the Partnerships for Advanced Computational Infrastructure (PACI) program, investing simultaneously in the National Computational Science Alliance, led by NCSA with partners at Boston University, University of Kentucky, the Ohio Supercomputer Center, the University of New Mexico, and University of Wisconsin; and in the National Partnership for Advanced Computational Infrastructure (NPACI), led by SDSC, with partner computing centers at Caltech, University of Michigan, and Texas Advanced Computing Center. An additional PACI award was also made for training and education, headed by the Center for Technology in Government at SUNY Albany with participation from both the Alliance and NPACI [1].

3.2 The TeraGrid

Following a President's Information Technology Advisory Committee (PITAC) report which had strong recommendations about needed investment in high-end computing acquisitions as well as research in high-performance computing [84], the NSF funded a program in support a of pushing national capabilities to the "terascale" range–that is, teraFLOP computing speeds, terabytes of data storage capacity, and gigabit networking links.

The first award of this program went to PSC in 2000 for the LeMieux 6 TeraFLOP system, followed by awards in each succeeding year through 2003 to continue building out terascale infrastructure: Distributed Terascale Facility (DTF) in 2001, Extensible Terascale Facility (ETF) in 2002, and Terascale Extensions in 2003. The name "TeraGrid" was given to the systems receiving the Distributed Terascale Facility award and applied to the extension awards in 2002 and 2003. The collaborative partiners for the 2001 DTF award were ANL, Caltech, NCSA, and SDSC. These four centers provided computational capacity at a collective 11 teraflops and 450 terabytes of data storage via a set of identical IBM linux-based clusters. Much of the initial expenditure of the TeraGrid awards was on connecting the sites with high-speed networks that would allow the computational facilities to be used in concert, as grid resources. The ETF award brought the PSC LeMieux system into the TeraGrid and augmented the capacity of the other TeraGrid sites. Terascale Extensions in 2003 brought in more centers to support the computational grid: Indiana University, Purdue University, Oak Ridge National Laboratory, and Texas Advanced Computing Center. In 2005, the NSF funded operations, user support, and further developments of the TeraGrid for a five-year period. During this time, additional awards added other sites as resources, starting with the National Center for Atmospheric Research in 2006, and the Louisiana Optical Network and National Institute for Computational Sciences in 2007 [172]. My own involvement with the TeraGrid began when I was assigned re-

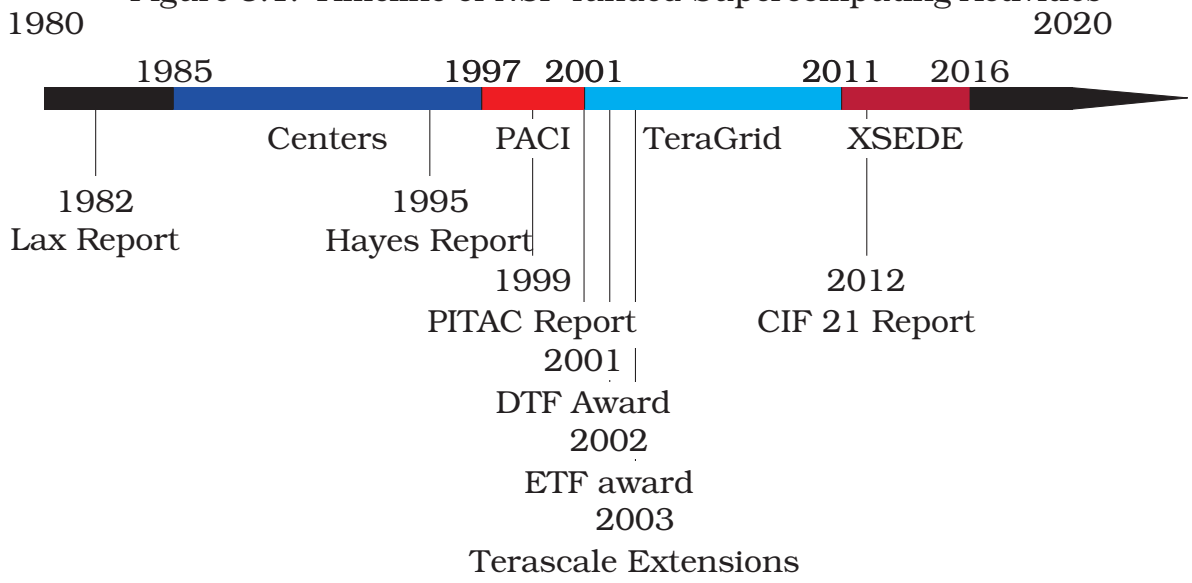
sponsibility for the TeraGrid accounts and accounting service for Indiana University's Big Red supercomputer system in 2007.

Following the end of the Terascale extensions award, the NSF opened a new solicitation for proposals for the successor organization to the TeraGrid with the Extreme Digital solicitation, which had its own repercussions in the community, as I describe in chapter 7, "Findings". Shortly after XSEDE began operations, in 2012, the NSF drafted a new report on major initiatives, entitled Cyberinfrastructure Framework for the 21st Century (CIF 21), which would allow coordination of the various directorates of the NSF and the cyberinfrastructure investments, and providing capabilities for science community-building, data-intensive science, and computational models. CIF 21 activities include prescriptions for research use of XSEDE but also guidance for individual projects which had their own considerable computational needs, and data-intensive activities that would support new analyses and greater interoperability between projects. A timeline of the projects, reports, and major awards is visible in Figure 3.1.

3.2.1 TeraGrid innovations

The institutions supporting the TeraGrid were funded in a number of "cross-referenced but independent awards to autonomous institutions" [40] arrangements which funded each of the individual institutions operating a supercomputer accessible to users of the TeraGrid, as well as

Figure 3.1: Timeline of NSF-funded Supercomputing Activities



institutions supporting the overall framework of the TeraGrid. The supercomputers and storage systems were referred to as *resources*, while the institutions operating these systems were *Resource Providers* (RPs). In addition to the RPs, the NSF funded the Grid Infrastructure Group (GIG), headed by University of Chicago/ANL, to provide direction in security, authentication, resource allocations, accounting, and overall TeraGrid operations including the help desk. The principal investigators at each of the RPs and the GIG made up the TeraGrid Forum, the center of strategic leadership for the TeraGrid and coordination among the RP's and with the GIG. TeraGrid Working Groups and Requirements Analysis Teams (RATs) provided guidance on broad, project-wide concerns and targeted, specific initiatives, respectively. In order to provide outside input to the TeraGrid

organization, the NSF mandated the creation of the Cyberinfrastructure User Advisory Committee (CUAC) [172, 40, 41].

As a large-scale distributed virtual organization, the TeraGrid dealt with the considerable challenges of making the first steps towards creating a common cyberinfrastructure. Accounts and accounting tools were developed and implemented, initial rules for a common unit of computing were developed and an allocations process created which would award users with units to be spent on systems. The GIG developed a considerable amount of software which would let users know the status of a given RP system and provided test harnesses for frequently-used software. Beyond the Common User Environment – software implemented for use by users of the supercomputer systems – the AMIE accounting software for distributing account information and monitoring usage [137] and the INCA test harness system for monitoring software availability [142] were important components of the grid infrastructure which ensured respectively, that user information could be distributed and synchronized, with common accounting of usage across many systems in the grid, and that systems which were expected to run a given software program could do that on a reliable basis.

As noted below, there were tensions between RPs and around the implementation of the Common User Environment. The NSF's set of "cross-referenced but independent awards" created a virtual organization which allowed for the establishment and growth of weak ties [75] between su-

percomputing centers and new organizations. In addition to expanding the homogeneity of the TeraGrid hardware and software systems to new and different systems throughout the life of the TeraGrid, as the membership of the TeraGrid expanded beyond the original set of Supercomputing Centers, more organizations with different cultures joined, bringing new sets of expertise, different perspectives on usage and policies for sharing resources, and with their own individual incentives and missions.

While these differences contributed to conflict over practices and activities on the TeraGrid, they also brought together staff members to participate in the Working Groups and RATs and increased the range and diversity of opinions and perspectives. Other activities conducted during the duration of the TeraGrid supported and extended these ties beyond the reach of the TeraGrid awards: a number of NSF workshops conducted by the Office of Cyberinfrastructure, the development of the Coalition for Academic Scientific Computing (CASC), and the development of regional partnerships that allowed for cooperation between different providers of cyberinfrastructure at national, regional, and campus levels. There were also initiatives within the TeraGrid, notably its Campus Champions program, which developed the community of cyberinfrastructure users and providers. At the same time as this culture of cyberinfrastructure professionals was taking shape, the advent of commodity cluster-based computing and the rapid uptake of computational methods by many in the natural sciences meant that membership in the cyberinfrastructure commu-

nity greatly increased, no longer restricted to a few highly-funded centers with staff around single "leadership" systems. Taken together, this period from 2001-2011 represented the formation of a broader community of cyberinfrastructure providers and users. Within the TeraGrid, staff commended the new arrangement for creating a network of like-minded application specialists, systems staff, and more [172] who could contribute to each others' work in infrastructure development.

While the arrangement of the TeraGrid by the NSF lacked the accountability of a traditional hierarchical, bureaucratic organization, it did provide the basis for creating a network of professionals who worked on similar projects, developed experience with one another. It also provided a fair test of the ideas put forth by Foster and Kessleman about a distributed grid that could provide computational capacity in the fashion that a public utility might. While the idea of on-demand computational supply is simple to imagine and an attractive solution to the computational needs of researchers, the TeraGrid with its mission to provide a unified infrastructure to any scientist at a U.S. academic or non-profit research institution, faced a great number of challenges in providing a secure and flexible set of resources. First and foremost, much like any utility, a large-scale distributed computational grid needs some way to keep track of all of its users, and to keep track of their utilization. This in turn required the specification of some kind of standard amount of usage. The distributed grid was never intended by the NSF to be a collection of homogenous sys-

tems – the nature of large-scale infrastructure is that it is built out of multiple disparate efforts, connected by common standards. In order to have a standardized way of counting resource use and reporting service delivery, the TeraGrid developed a normalized unit of computational time (based on one hour of CPU time on a Cray X-MP system). The normalized unit (NU) is scaled into the allocation measure of the TeraGrid, the service unit (SU), which at the time of the TeraGrid, was roughly 20 NU's to 1 SU on a contemporary processor for the purpose of allocations [41]. The standardized SU allows for acquisition and incorporation of increasingly powerful computers, while continuing to have a baseline "currency" for the allocation and usage of resources.

The allocation process was another necessity to having a successful distributed research infrastructure. As an alternative to a market for allocating resources, which, to be fair, presents a number of significant challenges in ther research context, the TeraGrid solicited applications for SU allocations, which would be awarded based on scientific merit and feasibility. Researchers applied to the TeraGrid for allocations based on particular scientific proposals, discussed their scaling studies to be sure that software would run and run efficiently on TeraGrid RP systems. The Resource Allocation Committee (RAC), made up of other computational scientists, serving for a 2 or 3-year terms, decided the viability of the proposal, and how much of the TeraGrid resource to allocate to that project. In some ways, it is telling that the TeraGrid should choose a method for

allocating resources that is fairly similar to the model of its parent organization, the NSF: there is a set amount of available resource, scientists propose to use some fraction of that resource, and other scientists determine whether the proposal has merit and the amount of resources to allocate. This means of allocation is in alignment with the funding organization's own processes, and the strategy would be well-understood by the NSF reviewers of the TeraGrid project itself. The RAC oversaw the selection of allocations proposals for requests more than 500,000 SU's (large requests) on a semi-annual basis, and requests between 30-500,000 SU's on a quarterly basis. In addition to these allocations, it was possible to request a "startup" allocation of less than 30 SU's, reviewed by TeraGrid staff, and commonly focused on testing and development activities for implementing scientific software on the TeraGrid. Having a successful startup allocation and demonstrating that software could scale to take advantage of large computational resources was frequently cited as a necessary component for a future larger allocations request.

Demand for computational resources made available by the TeraGrid were considerable, and the RAC was required to make determinations about a significant number of requests for resources. Catlett et al. [41] report that in 2005 and 2006, the requested NU's for the TeraGrid were 147% and 132% of the available capacity of about 881 million and 2.23 billion NU's, respectively. Allocations were awarded that were either specific to a given TeraGrid resource, or *roaming* allocations which could be

spent at any computational resource. In effect, the allocations process gave TeraGrid users a set amount of resource that they could spend down over the life of their project. Researcher NU's were not fully analogous to a "currency of the TeraGrid", however. NU's for specific allocations could not be converted into allocations on other resources. A researcher could create an extensive project staff which would be able to spend on the project's allocation, but researchers could not exchange NU's, for example, other than allowing other researchers to join their project and spend SU's. Researchers could see their consumption and would be alerted when they used more than the allotted amount. On some resources, they would not be allowed to submit new jobs without replenishing their allocation by submitting a renewal or extension. Analysis of usage by management of the TeraGrid took a fair amount of effort during the early years of the TeraGrid, as RP's collected job data by logs and aggregated it on a quarterly basis for the RP and later integrated reports to the NSF. Later development of the Metrics on Demand service [68] would provide a location and functions for aggregating and analyzing usage.

3.2.2 TeraGrid Metrics and Incentives

Any government-funded activity needs to show results for the expenditures it incurs, and the TeraGrid organization was no exception. Identifying the results of programs designed to support basic science with computational resources was not an easily solved problem. TeraGrid managed

the reporting responsibilities variously by tracking service delivery, outputs, and outcomes. Service delivery was the simplest task of these to tackle, as sites could track utilization and number of jobs, as well as system availability, to show that services were being delivered consistently and usage was supported. Outputs were tracked in the form of user publications. While users were encouraged to attribute TeraGrid resources in their work and report publications that were supported by the TeraGrid back to the organization, this was not incorporated into the allocations process as a requirement until the end of the TeraGrid awards. This lack of feedback from users resulted in TeraGrid being unable to easily record outcomes throughout most of its operating lifespan. Finally, outcomes were solicited in the form of scientific discoveries, which were made available to the NSF variously as "science highlights" or "science nuggets", which could be provided in press releases or to congress, with information linking the discoveries to TeraGrid activities.

These metrics for reporting, and the fact that initially, individual awards were expected to report individually resulted in a few difficulties and perverse incentives in the organization. As noted above, initially each RP and the GIG was required to report separately, only late in the life of the TeraGrid moving to a unified report structure guided by the TeraGrid Forum. The difficulty in providing homogeneous systems despite the creation of the Common User Environment, and the uneven application of software, paired with metrics per-RP, rather than per-project in the early days of

the TeraGrid governance structure, created an incentive for RP's to capture users, by providing better support for users that tended to use more resources, and making sure that they continued to be happy with the resources they were running on. This was aided by the fact that users tended to use and trust resources at their local institution, rather than those provided by other centers. Finally, the lack of simple mechanisms to transfer data around the TeraGrid contributed to the transition costs that users bore if they moved from one resource to another. Dan Reed, at that time at NCSA, was attributed to have coined the saying "Having people use your supercomputers a lot is very nice, but if you have their data, you have their soul." By retaining users who would consistently use a particular RP's systems, for instance by making sure they had the needed software for their analyses, providing better support, or by serving a sector that other resources did not, RPs could demonstrate their systems were necessary and could fulfill metrics on their own awards, ensuring that their demonstrated performance could be used as material for later proposals.

Other characteristics of TeraGrid rules and reporting made for conflicting incentives within the organization. Based on my interviews with staff involved with the TeraGrid project at Resource Providers, I learned that for a period within the program, officers at the NSF stated that only distributed work would count as TeraGrid jobs, that is, metrics for the TeraGrid meant that jobs which were executed on more than one sys-

tem counted as service delivery by the system. While this did provide an incentive to capitalize on the distributed nature of TeraGrid and foster development of grid computing, this requirement was put into place after the move from four identical systems to multiple homogeneous resources. In reality the software in place for managing workflows which span more than on system was still in the development phase. While demonstrations utilizing multiple systems were conducted, user workflows that made use of more than a single system were a rarity. As such, the TeraGrid Forum members regarded the requirement as a mismatch, obscuring the actual service delivery which the TeraGrid conducted, in favor of incentivizing the support of research which was in reality quite rare, and difficult to carry off. Another de facto requirement imposed was in the form of the NU for measuring utilization. My respondents also noted that the NU as originally designated in the TeraGrid strictly enforced the characteristics of systems which could be used as RP systems. There was no allocation equivalency for storage or visualization resources. Systems which utilized different architecture than standard benchmarking tools required, either as their central processing unit, or as what we now describe as accelerator cards, were not allocable resources, either. This approach meant that the traditional high-performance system as the main means of providing computational resources was reinforced, and new and innovative technologies would not be counted as contributions to the TeraGrid. While this cemented the notion that TeraGrid was a production science platform,

rather than a research and development project, it limited the flexibility of the TeraGrid significantly, and not having an allocable visualization resource also considerably reduced the range of offerings that TeraGrid could provide, despite the fact that most TeraGrid partners agreed that visualization was an important part of the research process. The lack of allocable units for storage cyberinfrastructure also meant that a TeraGrid file storage solution, whether distributed or centralized, had little weight in the considerations of the TeraGrid forum. Between the omission of storage metrics for service delivery and the incentives across RP's for user capture mentioned above, there was little impetus to include capabilities that would ease data transfer around the TeraGrid. These disincentives were part of organizational tensions Zimmerman and Finholtz identified and that I describe in the following section.

3.2.3 TeraGrid tensions and dynamics

While the goal was to create a distributed, interoperating system, the partners in the TeraGrid did not choose each other (except for a joint proposal between IU and Purdue), and the NSF managed growth with succeeding solicitations for resource providers, rather than providing a strategic blueprint for integration. One of the difficulties cited by the leadership of the TeraGrid during the Terascale Extensions period was in successfully managing strategically in an environment where additions to the TeraGrid distributed cyberinfrastructure were uncertain. With the goal of pro-

viding new and innovative cyberinfrastructure systems, the awards that funded hardware acquisitions at the RP sites were focused on large capacity, brand new systems implementing new technologies. The NSF provided solicitations for new cyberinfrastructure proposals with fairly broad scope, so that TeraGrid leadership was aware that new resources and new RP's were on the way, but it was not possible to predict what the new resources would look like or who would be providing them. A situation could arise where the CUAC requested a new functionality that the GIG analyzed and developed, which would be impossible to implement on a system awarded by the NSF the following year. This structure for adding new resources to the national cyberinfrastructure made planning for the future difficult for the TeraGrid GIG as well as the RP's [153].

Owing to its structure as a set of multiple-interrelated programs, the TeraGrid exhibited a number of tensions and dynamics. These tensions are embodied in some of TeraGrid's major charges: supporting deep research in the computational disciplines, broadening use of computational resources to new communities and new domains of research, and providing an open cyberinfrastructure for further extensible development. First and foremost, the structure of multiple RPs operating together with the GIG made for an uneasy relationship between the varying service delivery mechanisms of the project. The supercomputing centers which were TeraGrid RPs still carried out other activities for their local institutions, and still needed to compete for and win NSF awards for new research and

new infrastructure. This meant that the RPs operated in an environment that was collaborative and competitive by turns – the centers acting as RPs still needed to retain autonomy and show that they had advantages over other centers. Other institutions, who had not been part of the supercomputing centers program, had an incentive to win RP awards, or in some cases, provide local resources in order to improve their own standing as institutions capable of implementing and managing RP systems, and further support their activities as part of the national cyberinfrastructure community. Under the TeraGrid, RPs were not always incentivized to cooperate with each other. While the GIG was charged with providing a common infrastructure and operational software, as well as the Common User Environment (CUE), interests of the RPs were not necessarily served by installing all components or providing a common set of interactions for every TeraGrid user. This tension was also evident in relationships between local responsibilities and responsibilities to the TeraGrid virtual organization. Under the TeraGrid, as is common under most grant-funded activities, many staff had percentages of time dedicated to the TeraGrid while retaining additional local responsibilities. Local supervisory staff had little insight into what staff responsibilities to TeraGrid actually constituted, and TeraGrid staff reported that getting responses from a staff member at another RP site could be problematic. A collection of independent awards to different institutions, the TeraGrid had minimal controls for ensuring that all RPs were working in concert according to TeraGrid

goals [172].

Likewise management and coordination functions were not built into the project as part of the NSF's solicitation for the TeraGrid partners or RP's, forcing the TeraGrid members to develop their own strategies for unified project management. The resulting situation is referred to in one account as the "post-facto application of project management". While many project management processes are based upon the assumption that the project revolves around development of an idea, proposing the idea to stakeholders, and executing the construction or implementation of the project, the TeraGrid activities consisted of the operation of a distributed cyberinfrastructure (the services provided by the RP's) but also the process of identifying, developing, and integrating new capabilities (GIG responsibilities of analyzing needs, gathering requirements, creating solutions, and testing them). This required the TeraGrid to adapt management strategies that would assure support of operations in a reliable fashion as well as development activities that would meet the need for new software activities. The leadership of the TeraGrid evolved project management over the course of the Terascale Extensions award, from separate activities for the GIG and each of the RP sites, beginning to unify planning and reporting functions for all members in the third year of the project, and after some difficulty in identifing the correct scope for work breakdown structures and program plans, creating a fully-integrated planning process for all members in the fifth year of the project [153]. The adaptation

of the TeraGrid Forum to incorporate project-wide planning and management in order to deal with the difficulties of managing a large-scale distributed project highlights some of the challenges that make long-term cyberinfrastructure integrating multiple resources over the long term a complex process. Even for the fast-paced development of HPC systems, the lifetime of a system, and the need to interoperate over the long term, mean that infrastructure must be managed strategically, and network relationships must be effectively leveraged in order that the organization can continually deliver services and not find itself bogged down by internal difficulties. Nevertheless, the TeraGrid Forum could only make decisions that affected all members with complete consensus among those members. This *liberum veto* meant that any of the RP's involved in the TeraGrid Forum could effectively stop a new development.

A second set of tensions identified by Zimmerman and Finholt lay in the mandate of the NSF, continuing through the XSEDE project as well, to provide next-generation resources at the edge of technological capability, but also to engage new computational users from diverse communities, as well as to provide robust services for the broader CI environment. This put the TeraGrid in the position of needing to serve both highly advanced users as well as very inexperienced ones, including those who were encountering computational methods for the first time. Providing high-quality and available services serves both the advanced computational and new users well, but it put the project at odds with the goal

of providing next-generation resources, which by definition have an experimental nature. These tensions, when coupled with the RP model for providing services, meant that some RPs were better-positioned to provide the higher-capability services, and some were more focused on general-purpose computing that would be accessible to new communities. Certain of the TeraGrid resource offerings were aimed at particular groups of researchers: the Blacklight system at PSC shared up to 32 terabytes of system memory among all nodes of the system and remained in service for a number of years based on the needs of software for the genomics and natural language processing communities. Other systems, including those provided by new entrants to the cyberinfrastructure community, had less of a fit, based on the local site expertise (in contrast with the decades of experience at original NSF Supercomputing Centers program sites) or different architectures. Observers reported that some activities needed to be done to bring users to those resources and help them adopt [172]. In one case, local users at the resource provider's institution were advised to access the resource through the TeraGrid in order to help drive adoption of both the resource and the TeraGrid's general offerings.

Zimmerman and Finholt outline a third set of tensions in the TeraGrid project, that of the tensions between reliability and sustainability of infrastructure and the research and development function of implementing new, large-scale systems and ensuring interoperability with future systems and architectures. In their report, they note that staff and users

themselves had trouble defining the balance between TeraGrid as a research and development project aimed at improving computer science and TeraGrid as a cyberinfrastructure which would provide long-lived support that would be extended over time to new and different systems. While all of the staff interviewed for the report agreed that no computational research should suffer as the result of research and development activities in the TeraGrid itself, past that point, opinions differed greatly about the activities of the TeraGrid in relation to providing resources for science versus being a science project in and of itself. Certainly, at the start of the Distributed Terascale Facility, the emphasis was on creating a homogeneous grid system – the resources were homogeneous, and they were designed to act as a linked system via dedicated networking. Under the DTF, the systems offered by the four centers, their software and configuration could largely be managed in lockstep. With the introduction of the Pittsburgh systems under the ETF, the TeraGrid was now dealing with a heterogeneous system, which required considerably more coordination and also limited the ability of software, which was frequently compiled by users with their own configurations and tunings, to be copied to other TeraGrid systems and run without problems. With successive additions to the the project in ETF and Terascale Extensions, the goal of homogeneous distributed systems was discarded and the task of the TeraGrid was to create a flexible and open environment that could support computational usage across all sites. Differences between RP systems persisted,

however, and the difficulty of getting a particular set of analyses set up and running on the TeraGrid required an investment of time which meant that there was a considerable switching cost for all but the most flexible or design-oriented researchers. For their part, most researchers noted that they preferred to focus on the execution of their analyses in place, rather than making code portable to the multiple systems of the TeraGrid.

At the close of the TeraGrid project, the partner projects had created the basis of a distributed computing cyberinfrastructure. The GIG and CUAC were able to implement a system with allocated resources, based on peer-reviewed requests. A general parity of software packages was available throughout the TeraGrid, despite difficulties in getting all TeraGrid Forum members to agree. Researchers could take advantage of significant resources without spending funds for computing power or having to write expensive purchases into grant proposals.

Chapter 4

Research Methods

This investigation utilizes both quantitative and qualitative techniques in order to provide an informed picture of XSEDE and the national cyberinfrastructure community. As part of my responsibilities working for the XSEDE project as a deputy and then manager for the XSEDE Campus Bridging team from 2011-2016, I had considerable access to XSEDE management meetings, allocations meetings, staff meetings, and project activities involving XSEDE users. This access, with the support of the XSEDE Senior Management Team, afforded considerable activities for observing interactions among staff, including between staff of different supercomputing centers, as well as staff interactions with XSEDE users and between users themselves. In order to assess some of the claims being made about XSEDE's activities and those of its user base, I turn to the extensive amount of usage and publication data which XSEDE collects in order to measure and assess performance and report successes to the National Science Foundation and other stakeholders. This data provides me with some means to investigate the ideas laid out in the interview and

observational data. In the following sections I provide some detail on steps taken to ensure that my own observations are reliable and not unduly influenced by any party, describe the activities for taking data qualitatively and quantitatively, and what kinds of methods I used to assemble my quantitative findings.

4.1 Investigator Statement

As part of the explication of data collection and generating reflections on the data, it's necessary to describe what steps I have taken to maintain a perspective that is not overly influenced by my informants. Furthermore, as a professional IT worker who is paid to participate in the project that I am observing, I must make pains to identify influence and be aware of what the implications of that relationship are. Instead of being an investigator who is allowed access to the project and its staff in order to conduct interviews and observe, I have been a part of the project efforts since the project's beginnings in 2011 and through the completion of XSEDE, and currently in the follow- on project. While conducting these observations and interviews I was first a deputy and then manager of a project team working in XSEDE management. This afforded me a level of access quite beyond what a regular investigator might be provided. As a manager within the project structure, I participated in both staff and management meetings, was able to have conversations with the Principal Investigator of the project, as well as NSF Program Officers for the project, I participated

in the drafting of reports and proposals associated with the project, and had to formulate policy for my team's initiatives. I also was able to discuss my findings and reflections with the XSEDE external evaluation team. My situation with XSEDE and my research on the project represent a balance of access to informants and materials against becoming too close to the project mission, goals, and activities.

Part of this balance stems from the fact that, while I have worked in IT for a number of years, my background is far removed from that of most of the management and staff of the XSEDE project. Rather than being trained as a computational or computer scientist, my background is in the humanities and the social sciences. I came into the academic IT world shortly after I arrived at Indiana University to pursue my masters' degrees, working in a support role for unix systems in university departments. My own computational methods are limited to social network analyses and some work in basic statistics which require the use of supercomputing systems at their most rudimentary level. For the most part, I remained outside the world of NSF cyberinfrastructure until I quite nearly simultaneously started my PhD Informatics studies and started work with the XSEDE project. This conveniently provided me with a topic of study that I examine in many of my classwork activities. I would eventually develop a dissertation research project that would benefit from participating in work with XSEDE as a project team lead and working for the first time with other team members in a virtual organization. This was facilitated by

the XSEDE organization's openness to allowing a number of researchers to conduct research [86, 27, 172].

Not only am I not alone as a researcher of XSEDE, the distributed nature of the project and its membership means that the project staff of XSEDE and my informants come from a fairly broad spectrum of backgrounds. While I may be one of the few individuals involved in XSEDE who does not come from a background in the natural sciences, the members of the community that I encountered in interviews and discussions have a fair amount of diversity, albeit a diversity that is cultivated rather than organically representative. The XSEDE project takes pains to increase inclusivity and representativeness as part of addressing NSF's "broader impacts" issues. While the management of XSEDE is largely white and male, initiatives to increase the inclusivity of the project are in evidence. While selecting informants I attempted to engage not only those who had been in the project for long periods of time and were involved in the center of the organizational activities, but also those who had had marginal interactions and had thoughts about the nature of the project and its activities, services provided, and the common modalities of use.

Apart from coming to the community as an outsider of sorts, I have throughout the process of data collection, meeting with XSEDE staff and users, and observing the cyberinfrastructure community, attempted to provide members of the community with opportunity to comment and improve upon the findings. This provides the community with a means of

reflection on those findings and to provide their own voice in the research. In my observations of XSEDE and the national cyberinfrastructure community I have come to find plenty of differing opinions on the way things are and they way that they should be, but the vast majority of the XSEDE staff and management I've observed have been motivated by the support of science more than any other factor. As I have seen, the XSEDE community strives to be disspassionate, egalitarian, and inclusive, and it is my hope that this research project embodies those values as well.

4.2 Qualitative Analyses

The XSEDE project provides ample opportunities for qualitative data collection. As part of my activities with the XSEDE project, I was able to observe interactions between staff, management, and users. I was also provided access to XSEDE's documents (most of which are available to the public), but also was party to the drafting process of many of these documents, which greatly improved my ability to observe the conduct of XSEDE as the project created its narrative to the NSF, to its advisory boards, and to the researcher user base. Below I detail the forms of document analysis, interviews, and ethnographic observations conducted as well as the context for data collection.

4.2.1 Document Analysis

In order to understand the project activities for areas I was not immediately involved with, as well as to keep acquainted with the project structure in terms of who was responsible for what, as well as to maintain understanding of the organization as it went through a number of minor reorganizations throughout its five-year span. Throughout my work with XSEDE, I frequently turned to XSEDE's internal documentation in order to orient myself as to the project organizational structure, understand reporting lines, and to hypothesize relationships between units within XSEDE. Due to the changing nature of NSF requests for information, frequent reviews and suggestions on how the project might be improved, XSEDE tends to be a moving target. XSEDE project documents allowed me to understand what changes occurred over the life of the project (2011-2016) as well as to identify points at which some critical input had entered the project. For the most part, documents within the XSEDE project are available to the public as the result of government-funded scientific activities. The XSEDE Senior Management Team approved my use of XSEDE systems and documentation that supports daily processes and execution of the project's responsibilities.

The project maintains an extensive wiki with minutes of weekly team meetings, project meetings, and quarterly management meetings, as well as for project activities such as software architecture development, policy formation, and planning and executing team projects. The wiki is also

used to collect documentation in draft, provide a space for team members to review proposed initiatives, and to organize logistics. In addition to the informal documentation in the wiki, the project generates a significant number of public documents, for training and documenting procedures, on the XSEDE.org website. These public documents include the original XSEDE project summary and science case documents used in preparing the XSEDE proposal, quarterly and annual reports, program plans, the charter of the Service Provider Forum, staff climate survey, and evaluation reports.

In addition to reviewing the quarterly and annual reports and the XSEDE wiki, I also subscribed to XSEDE mailing lists and reviewed the contents of discussion. These lists included the Training, Education, and Outreach list, XSEDE Campus Champions list, and the Campus Bridging team list, as well as XSEDE team-wide communications and trouble tickets sent to my site as part of its role as a Service Provider. For user support, XSEDE maintains a trouble ticket system ("Request Tracker"), which captures help requests for the The XSEDE design, and which distributes trouble tickets to relevant parties within XSEDE as well as at the Service Provider sites. The development and implementation team and extended community support team make use of a Jira issue-tracking system in order to coordinate software development activites, which provided information on internal development efforts. Finally, XSEDE utilizes a Risk Registry system by which the project can track risks to its activities

and trigger responses to mitigate.

4.2.2 Interviews

Over the course of investigations of the project I conducted 22 interviews with XSEDE management and staff members, as well as members of the XSEDE Campus Champions group, who represent scientists at universities and bring new researchers to XSEDE resources. These interview respondents ranged from those who had been part of multiple centers throughout the Supercomputing Centers program, TeraGrid, and XSEDE to early career researchers at minority-serving institutions who were just starting to engage with computational sciences. In order to identify candidates I selected people affiliated with the XSEDE project who were central to the project such as the Principal Investigator and the NSF Program Officers (two program officers served during the project period), as well as the persons responsible for operations, software development and integration, allocations, and architecture. I also selected members of XSEDE staff who were involved in the areas of particular interest to me, including the leaders for broadening participation, for education and outreach, and external evaluation. For users, I selected users at Indiana University who were close by and engaged in using XSEDE systems, thanks to referrals from IU's Campus Champion, as well as a set of users referred to me by the leader for broadening participation. I selected campus CI providers from individuals who I had met in NSF workshops, XSEDE al-

locations meetings, and other project activities who were not necessarily "power players" in the field, but who I observed to be open and forthcoming in their discussions with other members of the cyberinfrastructure community.

Interviews with researchers, Campus CI providers, and XSEDE staff focused on how each user became involved with computational sciences, and how they started their interaction with the XSEDE project, the nature of their usage of XSEDE, and their use of different computational facilities provided at the campus level or other organizations. I focused on allowing each interviewee to elaborate on their own experiences with the project, their needs for computational support, and their interactions with resources and XSEDE staff. Interviews were conducted in-person (16 interviews), via Skype videoconferencing software (3 interviews), and over the phone (3 interviews), depending on the situation and availability of the interviewee. The in-person interviews were conducted at the XSEDE quarterly management meeting in August of 2016 (6 interviews), at the annual Supercomputing conference in Salt Lake City in November of 2016 (6 interviews), and in respondents' workplaces (4 interviews). Where possible, interviews were recorded with the permission of the respondent.

Interviews on average lasted just under two hours. Interview respondents were selected from management by soliciting from volunteers who found out about my project by my project presentation at XSEDE management meetings, by recommendation from other respondents, and by

directly recruiting central individuals, for example, the NSF program officers (2 interviews) and principal investigator (1 interview), members of the senior leadership team (5 interviews). The basic breakdown of interview respondents by role is detailed in Table 4.1. Other interview respondents were recruited via reference from the training and outreach and broader participation managers, as well as through contacts generated via a software pilot project conducted on multiple campuses through XSEDE. This pilot project was intended to test software that provided a distributed file system for access from XSEDE resources in order to improve the ease of data movement and job submission and required biweekly or monthly teleconferences with these program participants over the course of about two years. I also conducted "snowball sampling", that is each successive interviewee was asked to recommend further contacts in order to develop additional respondents. There was some overlap of roles in the respondents as described in the tables. Some of the respondents were generated via the XSEDE Campus Champions program, which recruits volunteers (both faculty and staff) on university campuses to provide training and outreach activities which foster the use of XSEDE resources, and as such have are in a sense both users of and involved with the workings of XSEDE. Another respondent is highly involved with the development of software which runs on a large number of HPC resources, including XSEDE resources, and who participates in XSEDE management discussions as well as in other cyberinfrastructure initiatives. Finally, NSF pro-

Category	Definition	Number of Interviewees
Users	University researchers who make use of XSEDE resources. May or may not have had an active allocation at the time of the interview	8
XSEDE Personnel	Individuals who are partially or fully committed to work on the XSEDE project	10
NSF Program Staff	NSF Program Officers for XSEDE	2

Table 4.1: Number and Types of Interview Respondents

gram officers were scientists who had made use of cyberinfrastructure in the past or had active research agendas which made use of cyberinfrastructure.

In selecting interview respondents I attempted to capture a broad range, from faculty who had not yet created XSEDE allocations but planned to make use of XSEDE resources in order to their own NSF-funded research to those who had been making use of XSEDE resources and developing software for grid infrastructure since the TeraGrid project. Staff were asked about their perceptions about the organization of the project,

tensions between local institutions and the larger organization, and the changing usage of XSEDE by new disciplines and new institutions. For some staff, I especially focused on their responsibilities in bringing new users to XSEDE and what those users needed in order to start making use of resources. Interviews with NSF program officers focused on the NSF's goals for the the project, the contrast between broader participation and next-generation computing, and the relationships between XSEDE partner organizations. Table 4.2 describes the respondents' demographic information, background, institution, and the interview format.

A list of sample questions from user interviews is presented in Appendix A. Each interview was based on these questions, and questions varied based on whether the respondent was a user (CI providers or campus champions were also asked user questions) or a staff member. I allowed respondents time to elaborate on their thoughts. Every interview started off with questions about how users got involved in computational research and with the XSEDE project. Users were asked about their use of XSEDE and other cyberinfrastructures, about their adoption of cloud resources and science gateways, and about the types of analyses they used these systems for, and the benefits they received from XSEDE. Staff were asked their understanding of the formation of the project as well as its individual directions and their own areas of responsibility within the project, and the types of user behaviors they observed as typical within the project.

Role	Gender	Race	Background	Interview Medium
user	female	African-American	Early career engineering faculty, Jackson State University	Skype
user	male	African-American	Tenured Condensed Matter Physics faculty, FL A&M	Skype
user	male	Middle-Eastern	Graduate Student, Indiana University	in-person
user, software developer	male	White	Director at Research Computing Center, University of Utah	in-person
campus champion	male	White	Campus administrator in Office of Research, FL Int'l University	phone
campus champion	female	White	Director at Research Computing Center, OK State University	in-person
campus champion	male	White	HPC Center Senior Staff, University of Arkansas	in-person
CI provider	male	White	Director of Research Computing Center, University of Miami	Skype
CI provider	male	White	Director at Research Computing Center, City University of NY	in-person
XSEDE staff/developer	male	Asian	Software Engineer, Indiana University	in-person
XSEDE manager	male	White	Research Center staff, National Center for Atmospheric Research	in-person
XSEDE manager	male	White	Software Engineer, Argonne National Lab, University of Chicago	in-person
XSEDE manager	female	African-American	Outreach Director, Southeaster Universities Research Association	in-person
XSEDE manager/site PI	male	White	Director at Research Computing Center, Indiana University	in-person
XSEDE manager	male	White	Software Engineer, Indiana University	in-person
XSEDE manager	female	White	Tenured Psychology faculty, Georgia Tech University	in-person
XSEDE director	male	White	Director at Research Computing Center, Cornell University	in-person
XSEDE director	female	White	Software Engineer, Texas Advanced Computing Center	in-person
XSEDE director	male	White	Systems Engineer, National Institute for Computational Sciences	in-person
XSEDE PI	male	White	Director at Research Computing Center, NCSA	in-person
NSF Program Officer	male	White	Tenured Theoretical Physicist faculty, NIST	in-person
NSF Program Officer	male	White	Tenured Engineering faculty, Purdue University	in-person

Table 4.2: Interview respondent demographics and background

For all of the interviews I kept a series notebook where I would write down notes immediately after the interview concluded, so as not to interrupt the flow of interviewing with writing. For interviews where it was feasible to record I used a digital voice recorder, a recording app on an Android tablet, or a Skype plugin that allows audio recording of Skype calls ("Skype Call Recorder" [11]). In order to get a feel for the course of the recorded interviews, I transcribed four of the interviews using F4transkript software [3].

4.2.3 Ethnographic Observations

I framed my engagement in participant observation largely as defined by Schensul, Schensul, and LeCompte [134] - that of learning by periodic working alongside participants in the setting of the organization. As part of my responsibilities for the XSEDE project I worked alongside XSEDE management and staff. XSEDE condcuts significant amount of its synchronous management and coordination functions through teleconferences. There are bi-weekly meetings that coordinate most of the levels of the XSEDE Work Breakdown Structure (WBS), including a senior management team meeting between the principal investigator and directors, level 2 directors with managers, and managers with staff members. Bernard and Gravlee [29] note that a certain amount of misdirection is required in observational studies of this type, and I found that my dual role as a person with responsibilities in the project, and someone new to

the project, meeting with many of the participants for the first time, I was able to couch my curiousity about project workings as that of new employee learning the ropes. My activities in observation often focused on following the discussion at hand, recording the interactions. For particularly controversial or complex interactions, I found that reflecting on it with another staff member (ideally one tangentially involved to the matter at hand) would provide additional perspectives on the matter. Following the guidance of other organizational ethnographers, I tried to engage with staff who could lead me to other engaged informants [169].

In my participation with the XSEDE project I engaged in teleconference meetings at all of these levels as well as with calls with the XSEDE Advisory Board, which is made up of volunteer scientists who provide guidance to the project on initiatives to support research. I also attended individual project meetings including a long-term software pilot project meeting to develop capabilities for sharing of data and compute jobs between campuses and XSEDE resources. While these activities conducted over teleconference were subject to the mediating effect of the technology involved, it is important also to understand that this is the daily context for most of the activities involved, and that for the bulk of XSEDE staff, in person meetings were restricted to the annual XSEDE conference or other large-scale meetings. Although the audio-only teleconference did restrict the richness of the material to be observed, there remained considerable signals to be examined, based on who attended, topics discussed, who

was paying attention, and what the individual actors had in mind.

Records based on participant observations were created with a basic format. I developed a habit of starting every meeting off by opening my notes to a new page and noting the date, meeting topic, and if any attendants were out of the ordinary, I would note that as well. I outlined the agenda of the meeting and recorded discussion points with the initials of the speaker in order to keep track of which staff made which suggestions. Nippert-Eng suggests a mix of diagramming, where helpful, in addition to text notes [115]. For my own observations, diagramming was not generally a feature of notes, perhaps due to the fact that many of the activities were conducted via teleconference, with no physical seating to chart or spatial relationship between participants. Certain activities might be captured in a flow diagram that would illustrate steps in a process, such as the convoluted sequence of steps to produce a quarterly report for the NSF, or the process for approving a piece of software for general use on XSEDE. For the most part I noted flow of activities in outline form, most commonly annotating with an arrow to indicate influence or effect and stars to indicate importance.

By dint of my management position in the XSEDE project, I was able to conduct considerable in-person participant observation of the XSEDE project's decisionmaking and coordination activities. Face-to-face interactions included meetings with the management team, the XSEDE annual conference, meetings and conferences focused on cyberinfrastruc-

ture, outreach and training activities, and XSEDE allocations meetings. Management meetings held by the XSEDE project are three-day in-person meetings in which project planning and coordination takes place, as well as airing of general issues and new courses of action. These meetings are also where changes in the NSF's requirements for reporting and XSEDE initiatives were communicated throughout the structure of XSEDE. The XSEDE annual conference is an opportunity for users of XSEDE and interested cyberinfrastructure providers to present and discuss novel uses of the systems, attend training and networking events, and to meet with other cyberinfrastructure users. During this time I attended NSF workshops and other group meetings that were also attended by management of the XSEDE project. These meetings included meetings of the Advanced Research Computing on Campuses (ARCC) group, Coalition for Academic Scientific Computing (CASC), Internet2, the annual Supercomputing Conference, and Open Science Grid All Hands meeting. While these functions were not specifically XSEDE-related meetings, members of XSEDE management would frequently attend these meetings and present on the work being done by XSEDE, taking questions and sometimes criticism on these topics. I also attended XSEDE functions to provide outreach and training to faculty at universities, where I and others presented on the available cyberinfrastructure, support, and services for their computational work. XSEDE allocations meetings were especially informative examples of interactions among computational users – these were meetings run by vol-

unteer users who reviewed and advised on the approval or rejection of allocation requests for service. While XSEDE staff facilitated the process and advised about new developments, determinations about allocations were up to the Allocations Committee.

Attending these meetings provided ample opportunities to see how staff from different centers worked with each other, and with NSF program officers for the project, both in the context of presenting the organization to users, and "behind the scenes" with staff members outside of open forums. My own position as a person new to national cyberinfrastucture and new to the organization afforded ample opportunities for me to attempt to fit in with these long-time cyberinfrastructure providers and users, and identify my own presumptions about the organization and how it operates, as well as see the stories these groups tell themselves about the development of cyberinfrastructure and computational research.

4.3 Quantitative Analyses

In order to further my examination of XSEDE activities and to attempt to quantify the linkages between service and science outputs, I examine usage data based on XSEDE projects and self-submitted publication data provided by XSEDE users based on resources used provided by the XSEDE project.

4.3.1 Data Gathering

For the quantitative portion of this research, two main types of data are used: XSEDE publications collected by the XSEDE user portal and provided by the XSEDE project, and XSEDE user records and project data, provided by the XSEDE Metrics on Demand project (XDMoD). The data provided by XSEDE covers a span from the inception of the Teragrid Central Database in 2003 through the transition to XSEDE in 2011 and up to July of 2016, when the XSEDE award completed. The data described below has been provided by XSEDE staff who create and present metrics for usage and application data, as well as by those who are engaged in project management and documentation of project results to the NSF.

Publications Data

XSEDE staff provided the contents of the XSEDE publications database, which is a self-reported database of publications supported by XSEDE. Individual users of XSEDE record their publications via the XSEDE user portal, where they can be viewed as part of the online user profiles, and also used by XSEDE in order to measure and demonstrate the scientific output of researchers making use of XSEDE resources. I obtained a dump of the XSEDE publications database in CSV format which included all of the publications recorded in the XUP since its beginning and ending July 1 of 2016. The publications database as provided contains 7981 submissions, of which 6883 are associated with XSEDE projects and 1098 are

recorded as supporting materials for XSEDE allocations requests, which XSEDE staff note are not the result of work done on XSEDE, but rather work preliminary to a request for allocations on the XSEDE project, and these records were removed for the utilization analyses as they do not represent utilization of XSEDE resources. XSEDE publications data, because it is self-recorded by the authors and not normalized by the XSEDE user portal's intake process, tends to be somewhat messy, and requires some processing. Journal author names, as well as publication names were transcribed to the ASCII character set from UTF-8. Some particularly long publication names included line breaks that needed to be removed from the initial data set in order to parse properly. For the purposes of a co-authorship network analysis, the data was reformatted as a bibtex file and author names were unified. Records that were not able to be parsed from the XSEDE data into bibtex were discarded. In all, 7978 total publications were obtained.

Usage Data

The original dataset on projects and users in TeraGrid was compiled with the assistance of the XSEDE accounts management team. A representative of the accounts team ran SQL queries against the XSEDE Central Database (XDCDB), originally the TeraGrid Central Database (TGCDB), which tracks all XSEDE accounts, allocations, and resource usage. The retrieved data covers all user and project information from the incep-

tion of the accounting system in 2003 through August of 2015. It includes information for 20,003 resource allocations, comprising a total of 28,567,137,013 Normalized CPU Hours, for 5352 Principal Investigators. XDCDB is populated by information collected by the Account Management Information Exchange (AMIE) system. All data was provided in comma-separated value files that can be easily processed programmatically.

The project data includes:

- Allocation short name, or Project ID and allocation identifier
- ID and name of Principal Investigator
- ID and name of the PI Organization
- ID, organization, and name of the XSEDE resource used in the allocation
- Field of science, identified from 147 specified fields
- Base allocation, the initial project allocation in service units (allocations can be extended for long-term projects)
- CPU hour usage of the allocation

Additional data was provided by the the XSEDE Metrics on Demand (XDMoD) site (https://xdmod.ccr.buffalo.edu). XDMoD is developed at the University at Buffalo and is detailed in [68]. It leverages the XDCDB as well as a number of probes which examine the performance of

XSEDE resources, including individual application performance. XDMoD includes information which can be explored by a number of means, including graphical display of usage by PI, PI Institution, Field of Science, and Allocation, among many others. XDMoD also provides a number of means for organizing and visualizing data about XSEDE. Data from XDMoD can be exported into tractable data formats such as csv, for programmatic manipulation. XDMoD staff provided support in querying and using the XDMoD database. Reports from the XDMoD database allow the aggregation of usage and allocation on a per-project or per-PI basis. Sample data from the XDMoD project is show in Table 4.3.

Allocation Name	PI ID	Resource	Field ID	Usage
TG-PHY100033	582	stampede	21	101611476
TG-MCA93S002	7	kraken	17	99563180
TG-MCA93S028	8	stampede	65	97955353
TG-CTS090004	7459	stampede	125	955559740
TG-MCB070015N	4801	comet	64	90510584

Table 4.3: Usage Data from XDMOD

4.3.2 Bibliometric Analysis

Using the XSEDE publication data, a co-authorship network was extracted and author names were unified using the Sci2 Tool described in [37]. The resulting co-authorship network has 11,063 author nodes and 32,048 collaboration links. This network was then analyzed with the MST Pathfinder algorithm in order to detect the backbone structure of the network. Weak component analysis was run to identify the largest fully

connected component. This resulting graph was analyzed for modularity in Gephi and color was used to indicate what authors belong to what cluster modules.

4.3.3 Analysis of usage data

In order to better understand the distribution of resource usage by fields of science within XSEDE, the project allocation data was aggregated by field of science. Fields of science were grouped into categories

To create maps of XSEDE resource consumption, institution data was matched with a lookup table of latitudes and longitudes provided by XSEDE. There are a few projects, such as the OpenWorm Project, which are virtual organizations that list no location. Exactly four organizations had a latitude and longitude of 0,0 and they were removed from the dataset.

Usage data is generated by user interactions with XSEDE resources and accounting takes place directly based on accounts used to authenticate and is tied to the XDCDB information, institutional and PI data is understood to be largely correct. The only instance of incorrect information included in this information would be if a user was using another user's account (a violation of XSEDE policies) or if incorrect information was entered into XDCDB.

In order to analyze the usage and publication data in respect to location, the Sci2 Tool was used to read the projects file and extract a two-mode network. The two-mode network has two types of nodes: re-

sources and organizations. Resource nodes have attributes such as location (lat-lon) and capacity (teraflops). Organization nodes have attributes for location and number of publications (aggregated for all users at an individual organization). Organization location is derived from XSEDE's table of all organizations that use XSEDE. The edges between resources and organizations are weighted by the size of the allocation in CPU usage. The resulting network was analyzed for centrality and node degree distribution using the Network Analysis Toolkit in the Sci2 Tool. Edges above 25M CPU Hours of utilization were extracted from the network and nodes were geolocated by institutional information, and the resulting network overlaid on a map of the United States.

4.4 Institutional Review

This research project has been approved by the Indiana University Institutional Review Board under Protocol #1605848427. Preliminary work for this research was conducted under Indiana University IRB Protocol #1505700642. Interview questions and study information sheets were developed with the help of the Indiana University Bloomington Institutional Review Board. The focus and aims of the research project was reviewed with the XSEDE PI and NSF Program Officer. For individual interviews and observation of closed meetings, study information sheets were made available for respondent/informant review.

Chapter 5

Findings

In order to describe my findings based on the conversations and interactions I have had with XSEDE users, staff, and management over the past few years, I will trace the organization from its roots up, starting with findings from XSEDE users, moving to groups of users (by organization, field of science), and then into the XSEDE organization itself, the interactions between the participating centers, and finally some insight into the NSF itself, as it relates to the provision of cyberinfrastructure support for basic science. Where it is applicable, I provide analysis of usage, fields of science, and publications in order to help inform the picture of XSEDE I describe. First, however, I will detail the transition from TeraGrid to XSEDE as related to me by a number of informants, in order to detail the structure of the XSEDE organization, the NSF motivations which defined XSEDE's mission and initiatives, and the relationships between the centers which shaped the organizational structure.

Following this first section on the transition between the two projects and the structure and makeup of the XSEDE project, I will detail my

findings about how individual users interact with XSEDE, how they work with XSEDE and NSF to get both access to computational resources as well as other benefits, and attempt to understand the makeup of XSEDE usage, by looking at the utilization of XSEDE by various fields of science and organizations. Following this, I examine the activities of the XSEDE project, describing the challenges and changes that my respondents working within the project described in their responses. While I attempt to attribute particular statements to individuals where this is helpful, much of what I report is the result of observation of conversations between staff, or informal conversations between users, based on my notes, taken either during meetings and staff activities or immediately after. In part this description is an attempt to present here the stories that the TeraGrid and XSEDE projects tell themselves, and make a note of where these stories differ from what I've observed on my own.

5.1 From TeraGrid to XSEDE

5.1.1 The XD solicitation

By the end of the TeraGrid project, the centers involved had fully completed the transition from a research and development project focused on the creation of a distributed grid-like system that could be used by a broad range of computational scientists to an operational cyberinfrastructure with heterogeneous members, focused on service delivery. The allo-

cations framework and development of the common Service Unit allowed for allocations and reporting processes to work smoothly. Many of the issues around this software had been brought to rough consensus over the course of the TeraGrid, but there remained issues with project governance and integration of Resource Providers. Furthermore, software quality for delivery to the resources was also the subject of much tension. As such, in specifying the activities of the project which would come after TeraGrid, the NSF focused largely on the quality of service delivery and the organization's responsiveness to user needs, rather than adoption of advanced technologies. As such, in June of 2008, the NSF released solicitation 08-571, "TeraGrid Phase III: eXtreme Digital Resources for Science and Engineering (XD)". The XD solicitation specified a set of key attributes for the distributed cyberinfrastructure which would succeed the TeraGrid. The solicitation focused on the development of the cyberinfrastructure based upon "sound system engineering principles", including a platform where the XD operations team would be able to test software implementations before releasing them to resource provider sites. The new organization would also be driven by the needs of its constituency, with an architecture team responsible for choosing new software capabilities based upon demonstrated user requirements.

At the same time, these key attributes required that the successor organization would implement existing software solutions, rather than developing new software to meet these needs. The key attributes re-

quired the organization to address a broad range of usage modalities. The proposers would need to provide support to researchers in need of sustained long-term usage of high-performance computing facilities and those who needed brief jobs run on fast systems in order to enable the interactive analysis and exploration of data. The new facility would also be able to support computations with minimal data movement as well as "data-intensive" ones. The XD award specified a set of services which the proposers would need to provide. The NSF described the intiatives as: *Resources* and *Integrative Services*. Resources consisted of computing and storage services which would be funded through the NSF "Track 2" program, and a remote visualization and data analysis service. Integrative Services would cover a number of functions required to support the cyberinfrastructure. A Coordination and Management award would provide the operations and security of XD. The "Technology Audit and Insertion Service" would identify potential technologies for the improvement and extension of existing computational capabilities. Advanced User Support Services would provide extended consulting services for adapting and optimizing codes to XD architectures, utilizing accelerator technologies, and supporting and extending science gateways which leverage XD resources. A Training, Education, and Outreach Service would provide training and enlist participation in computational sciences from a broad range of under-represented demographic groups.

The NSF specified that proposals for the visualization/analytics and

technology audit and insertion services would be separated from other proposals. Institutions could propose either an integrative service which would provide the coordination and management functions as well as one or more the other services described, or individual items from the integrated services. The operations would constitute the largest portion of the award, but with both the integrative activities and the computational resources part of the XD solicitation, coordination and accountability of the resulting organization would be much more clear than under the TeraGrid. The various institutions involved in the TeraGrid and some others quickly created alliances that would draft proposals to the XD solicitation's coordinating function, coalescing into two teams: the XSEDE team, composed of NCSA, the National Institute for Computational Sciences (University of Tennessee's part of Oak Ridge National Labs), Texas Advanced Computing Center, and Pittsburgh Supercomputing Center. The XRoads team was organized between San Diego Supercomputing Center, Indiana University, Purdue University, and Argonne National Lab. XSEDE and XRoads both submitted proposals for the coordination and management, advanced user support, and training and outreach activities.

5.1.2 The "shotgun wedding"

After reviewing the initial proposals and conferring internally, a process which took nearly a year, the NSF recommended that the XROADS and XD teams submit a new proposal, based on a combination of the proposed ac-

tivities. The resulting XSEDE team proposal incorporated members from all of the proposing partners, and has been referred to by some of the partners as a "shotgun wedding" between previous competitors. The new organization included both long-term centers with considerable user bases, as well as a number of smaller partners which had demonstrated capability in areas outside of the central service provision of XSEDE. Throughout this process, partners and TeraGrid RP's were expected to maintain teragrid operations without disturbing the research carried out on the systems, or the supporting processes like quarterly allocations of resources. In the following few pages I detail the changes to XSEDE from the TeraGrid project architecture and also describe some of the items that resulted from the incorporation of XROADS into XSEDE. Significant details from this integration process are available in the XSEDE Revision Narrative, which explains the response to the NSF's request to join the two proposal teams [167]. Meanwhile, other portions of the XD solicitations were selected with a minimal amount of controversy. The well-developed XD Metrics on Demand (XDMoD) service from the University of Buffalo under the leadership of Tom Furlani had analytics and visualization software available before the solicitation was released and stood to provide significant instrumentation to the existing resources. Pittsburgh proposed and was awarded for the Technology Audit and Insertion Service, based in part on the strengths of the software engineering strengths of the Software Engineering Institute at Carnegie Mellon.

Within the newly minted collaborative XSEDE project, however, the partners needed to achieve an organizational stability sufficient to allow them to pursue the objectives set before them. One of the more difficult issues was dealing with software architecture. The new XSEDE proposal included an architectural team incorporating two different technological capabilities developed during the TeraGrid years: the Globus project, headed by Ian Foster at Argonne National Lab, and the Genesis II based on Legion, a project headed by Andrew Grimshaw at the University of Virginia. Both Globus and Genesis II were software implementations that managed job submission and execution as well as providing data access and movement. Globus was part of the CTSS Remote Capability kit provided by TeraGrid RP's, but many found fault with the software, citing that the software was too difficult for most users to adopt, given the complex command-line structure, security certificates, and data management URLs involved. Genesis II was intended to make job execution easier and make use of remote filesystems to provide simplified interfaces for users to adopt, adapting job submission to a menu-based process that generated a submission file, including transfer of needed data into the compute systems and retrieval of results to the researcher's computer. Around the time of the transition between TeraGrid and XSEDE, Globus was undergoing an effort to reposition itself with more user-friendly services grounded in Web 2.0 design principles. The Genesis II project staff encountered struggles with sustaining sufficient adoption to get truly useful feedback

on the software, and maintaining sufficient development staff to act on design initiatives and provide a modern set of tools and interfaces.

Most of the parties involved saw the funding of XSEDE as a matter of survival for the relevancy of the associated centers. NCSA, SDSC, and PSC were original centers that had weathered significant successes and setbacks over the years. Argonne as a long-time center and the former seat of the TeraGrid GIG had always played an integral role within the cyberinfrastructure environment. Other partners, notably TACC, NICS, and IU, had developed significant capacity for acquiring and managing cyberinfrastructure resources during their parts in the Extended Terascale Facility. The losing team of collaborators might quite likely result in significant reduction in funding for new initiatives by the NSF and potentially viability. As the proposal process wore on over the course of multiple months, the prospect of restructuring of the community became disruptive, as anticipated changes to funding and structure resulted in staff leaving centers for other institutions, or leaving the research cyberinfrastructure community entirely. Informants from the project conjectured that the instability resulted in an overall reduction in available cyberinfrastructure staff, and certainly the number of staff members I encountered who had moved away from SDSC during this time to jobs at other partner sites seems to corroborate the assertion that those who felt that their livelihood might be affected by shifts in the funding environment took steps to find other positions.

Not all of the changes due to the shotgun wedding were negative. The creation of a more formal organization benefitted from the weak ties established under the TeraGrid days, as the makeup of XSEDE groups became more diverse in terms of which center provided the staff. Rather than remaining a largely University of Chicago group (from its roots in the TeraGrid GIG), the Software Development and Implementation team became a mixed team with membership from multiple centers, as occurred in multiple XSEDE teams. Rather than carving up the project into domains served by particular centers, the leadership of XSEDE approached the issue of filling each of the responsibilities of the organization by identifying the most appropriate staff from all of the partner sites in order to contribute to the different XSEDE functions. Most of the teams in the resulting organization were made up of members from across the partner institutions.

5.1.3 XSEDE operations begin: the move towards service

While XSEDE did face a number of challenges that were posed in the TeraGrid, and a few new ones due to its restructuring, the main focus of the project was first and foremost uninterrupted service delivery to its scientist constituents. Several of the systems developed under the TeraGrid, notably the AMIE accounts management systems and the RAS resource allocation system, continued to function as always. TeraGrid RPs at the end of the award period became *Service Providers* (SPs) under the XSEDE

award. The XSEDE help desk and support structure transitioned to a new ticket-tracking software, but the change was largely transparent to users and welcomed by staff, although considerable effort was taken to acquaint both with the new organization and state of affairs.

Formalizing the organization

Structurally, the new organization was a complete change from the organization of the TeraGrid, as already noted: management and reporting functions were organized under a completely top-down reporting structure. Under the advice of NSF program officers, XSEDE adopted software engineering practices intended to make sure that software provided to XSEDE SPs would be of high quality with sufficient implementation instructions that service provider staff could easily adopt and make new software and services available. In addition, funding the XSEDE integrating functions under a single award to UIUC, the leading institution, provided considerable structure to the organization that was not built in to the TeraGrid. The XSEDE project, once awarded, required a Project Execution Plan which described the overall functioning, requirements and deliverables, governance, and schedule for the project. Furthermore, the structure of XSEDE was designed around the idea that a central organization would provide operations and general functions that cross-cut the organization. The relationship between XSEDE and the resources it provides was made more flexible as well. Timing for the grants that fund

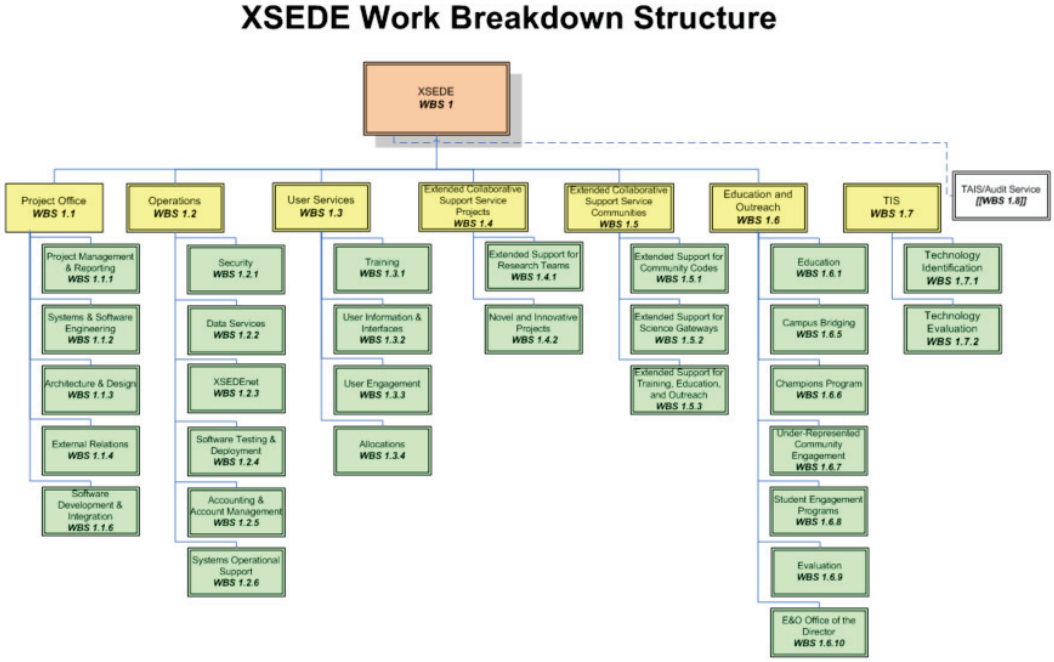
the acquisition of large resources and projects seldom line up easily. The XSEDE group created an architecture in which a Service Provider could create a resource aligned with XSEDE by installing a base set of XSEDE software, implementing security and authentication protocols that were accepted by XSEDE, and using the XSEDE accounts and usage systems to report usage against allocations. SP resources could come and go on their own timelines, and the XSEDE would continue to provide essential services for the operation of the distributed infrastructure. SP activities are governed by the SP Forum, in which issues between XSEDE and the SPs can be hammered out. The SP Forum has been extended under XSEDE to incorporate three tiers of Service Provider, based on the level of interoperability between XSEDE and the resource. This extensible framework allows for XSEDE to present a much greater range of resources to its users than if it were only able to offer systems created under NSF Track 2 awards.

As part of the strategy put forth in the Project Execution Plan, to introduce greater organizational clarity and accountability, the XSEDE project organized along the basis of some fairly common project management principles. The Work Breakdown Structure, or WBS, broke the XSEDE project into three levels with the PI, John Towns, as the Level 1, six Level 2 areas, and twenty-four Level 3 areas. Technology Audit and Insertion, awarded under a separate award in the XD solicitiation represented a seventh Level 2 area. Figure 5.1 shows the XSEDE Work Breakdown

Structure. Each of the Level 3 areas was represented as a team, with as many as ten or more staff members, or as few as one, depending on the area. Each of the WBS levels represents a hierarchical reporting line, with L3 managers responsible for staff and reporting to L2 directors. The L2 directors and PI Towns formed the Senior Management Team (SMT), which managed much of the tasks of collating and unifying report materials, responding to NSF requests, and coordinating cross-project activities. While the WBS allowed for accountability within the organization, some staff might be responsible for XSEDE activities that their local supervisor might not be involved in. In order to ensure cross-organizational accountability, XSEDE identifies local supervisors for every staff member funded at a partner site via the Statement of Work documentation for that partner site.

Another strategy that XSEDE incorporated in order to facilitate cross-site governance was the "canonical full-time equivalent" (canonical FTE). Staff salaries across fifteen different partners made for significant complications in organizing budgets. The canonical FTE would set a pay rate at which the project would pay partner institutions the same amount for any FTE working on the project – from PI John Towns to staff members – regardless of location or rank. The canonical FTE was $200,000, "fully loaded", meaning including salary, benefits, and facilities and administration for the staff member. This had the benefit of making centralized accounting and budgeting decisions simple enough to deal with more than

Figure 5.1: The XSEDE WBS (from the XSEDE wiki)



100 of XSEDE's staff. The canonical FTE also allowed for a direct correlation to be drawn between dollars and effort on the project, in a way that simple dollar amounts would not. While the canonical FTE did simplify counting on the part of the central institution at UIUC, and it did represent a way of equalizing staff commitments across organizations, it also put centers in areas with a higher cost of living at a distinct disadvantage compared to those in less expensive regions. Participant organizations had less incentive to put highly compensated staff on XSEDE roles, as the canonical FTE ended up paying for only a portion of these staff members, so those centers with higher average salaries had less incentive to participate, and all centers had less of an incentive to put highly paid,

expert staff on XSEDE. Whether this affected XSEDE's activities and service delivery is a matter of interpretation. When I have observed XSEDE teams selecting staff for specific activities, generally expertise is first and foremost on people's minds, and location (and salary) tends not to enter into the discussion. However, during my observations I have noted that program management staff, that is, those staff who are general project managers who assist with compiling reports and drafting planning documents, with less direct technical expertise, change frequently, sometimes from quarter to quarter. The conclusion I draw from this is that technical skills tend to be narrow, and individual XSEDE technical staff commonly have a niche that they fulfill (operations, virtualiation, science gateways), with a reputation for operating in that area, but that project management is a more fluid set of skills, and the project partners reallocate more frequently this work based on a number of factors.

User-driven requirements

As XSEDE moved into fully operational status, it began to adapt to requests from NSF reviewers and from the XSEDE Advisory Board, as well as from the Program Officer, Barry Schneider. In addition to providing frameworks for improved accountability within the organization and easing the transparency of both responsibilities and dollars between project sites, XSEDE engaged in activites to ensure that the XSEDE software would meet two criteria. Firstly, software and capabilities adopted by

XSEDE and made available to users would be driven by user requirements. Some discussion in response to TeraGrid activities focused around the development of software that was of interest to computer scientists and systems architecture rather than supporting scientific software. New software and capabilities (such as third-party data transfers or single sign-on authentication schemas) were to be driven by demand from the user community, rather than identified from within the project. The hope for this initiative was that there would be less time spent on implementing software configurations that were novel, but would go unused in favor of those that had user requirements.

As such, considerable weight was given to defining a set of use cases that would drive development within XSEDE. The use cases were fairly detailed documents. Secondly, the software delivered to XSEDE SPs would need to be operationally ready, meeting a set of quality attributes defined in the aforementioned use cases. As a result, the Architecture and Design group were charged with documenting the user needs driving new functionalities for XSEDE, and in some cases, documenting existing XSEDE capabilities in order to make clear that they answered user needs. The process of documenting need was a slow one, and the A&D team quickly found itself with a considerable backlog of use cases for software improvements to the TeraGrid, including painstakingly creating documents for capabilities users were already incorporating into their XSEDE work.

The work of the A&D team was also somewhat complicated by the gen-

eration of designs incorporating Globus, Genesis II, and the UNICORE technologies developed at ZTH Jülich. Some of the most contentious discussions I observed arose out of trying to meet NSF suggestions about methodology and architectural approach. The A&D team did provide a set of system-wide documents that describe the basic functionality for the XSEDE activities, in the form of the *XSEDE Architectural Overview*, which described XSEDE basic architecture functions as: identity management, interactive login, remote file access, submission and managment of computation, data transfer, and discovery and provisioning of resource information.

These basic functions are facilitated by three layers: access (interfaces for the user), service (connection to resources via standard protocols), and resource (the resources made available by the SPs) [151]. Guidance from the NSF on architecture decisions changed multiple times throughout the course of XSEDE, the team trying to variously meet requests for new functionalities that included built-in security and reliability for full production, later being directed to not focus on new development, but only on selecting and implementing cyberinfrastructure software developed by other initiatives. If the TeraGrid represented the NSF's first steps in moving distributed computing from a research and development project towards a production distributed computing environment, the XSEDE project marked the creation of a service organization which was concerned with not simply delivering hours of computer time and available systems,

but with the quality of service, its capabilities for new means of access and new activities, and with approaching different types of researchers to bring them into its user base.

Performance measures

In addition to newly-formed architectural processes for XSEDE, another evidence of the quality-of-service approaches of the organization arrived in the form of the adoption of performance management processes. In order to document and improve XSEDE activities, the organization adopted elements of the NIST Baldridge Performance Excellence Program [82]. As part of the Baldridge criteria adopted by XSEDE, activities included the identification of vision, mission, and goals statements, the identification of Key Performance Indicators (KPIs) supported by WBS area metrics. Each element of the WBS produced mission and vision statements that were to guide the activities of the group throughout the process. A significant amount of debate was spent on the construction of KPIs for each area and the metrics that would relate to them. As is common for organizations which have not constructed such metrics before, significant time was spent positing, adjusting, and learning about what activities could be counted. KPIs for a number of teams were revised multiple times throughout XSEDE in order to better capture the activities of the team and the outcomes desired. These performance management activities were augmented by several other activities, including the architectural process im-

provements related above, which were recommended by the Software Engineering Institute. Other activities included internal suggestions from area managers on process improvement, adaptations stemming from the cyberinfrastructure environment, and the staff climate survey. As part of the XSEDE Training, Education, and Outreach WBS area described below, XSEDE included an External Evaluation Team, which, with the help of other staff and area managers, constructed a number of surveys of XSEDE users as well as the staff climate survey, which allowed XSEDE staff an opportunity to suggest areas for improvement within the project anonomously, and particularly brought to light issues with diversity and gender relations within the project. The staff climate survey, conducted on an annual basis with voluntary responses from across XSEDE staff also examined relationships between different components of XSEDE, noting that communication and collaboration were seen as good, but needing improvement and standardization across the organization, and noting that transparency of the Strategic Management Team's decisionmaking process as well as that of the User Requirements Evaluation and Prioritization working group would greatly improve organizational alignment.

Another part of XSEDE which created formal initiatives built on ideas created in the TeraGrid was the Training, Education, and Outreach integrating activity, incorporated as an L2 directorate within the XSEDE WBS. Under TeraGrid, training and outreach activities were organized to recruit and train new users and provide students opportunities to work

with advanced systems. The XSEDE TEOS group provides user education, student engagement, the XSEDE "Campus Champions" program, in which volunteers at higher education institutions provide local outreach to researchers and help them make use of XSEDE, "campus bridging", which promotes the use of campus cyberinfrastructure to interoperate with XSEDE and ease the transition from campus to national cyberinfrastructure, and Underrepresented Community Engagement, which reaches out specifically to minority-serving institutions (MSI's), historically black colleges and universities (HBCU's), and tribal colleges. The overall mission of the TEOS group is to stimulate the adoption of XSEDE by new users, whether in traditionally computational disciplines or those new to the use of computational resources. In addition to providing next-generation computational services, the cyberinfrastructure provided by XSEDE can also provide basic resources for those at institutions that would not have the capital or technical know-how to provide their own cyberinfrastructure. What outreach staff for TeraGrid, XSEDE, and elsewhere has understood for a number of years about the difficulty for many at those institutions, however, is that there is a considerable gap in technical skills and computational understanding that needs to be bridged before researchers can make effective use of these systems. These difficulties remain in place throughout the XSEDE project, and they become more marked as the modalities we associate with personal computing become more simplified and more responsive, while the basic means of using HPC resources is the

same as it was during the Supercomputing Centers program - submitting a job to a batch processing system and waiting for results. While training and outreach activities provide a means to help researchers bridge that gap, as our day-to-day computing continues to diverge from computational research, that gap will continue to widen.

5.2 Understanding XSEDE users

In the following sections, I return to my research questions about the XSEDE user base and the XSEDE organization, starting with the user base, and discuss some of the claims that are made about the direction of computational users and the kinds of activities in which they engage. These questions, again, are *Who is XSEDE's user base? What are their needs? How do they get what they need from the organization?*

5.2.1 Changing fields of science in XSEDE

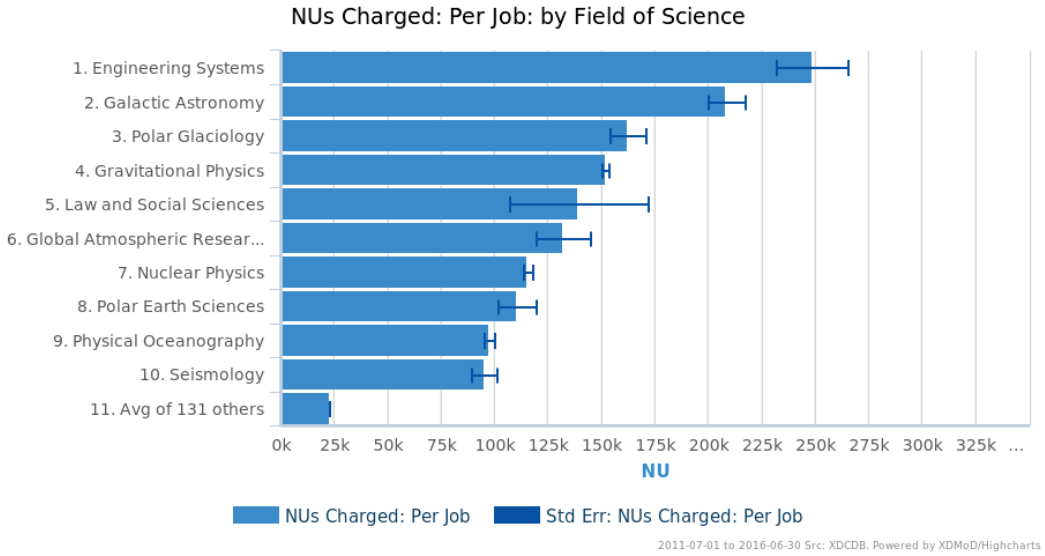
The XSEDE user base is encompassingly broad, and despite the challenges of outreach, there are a significant number of researchers making use of XSEDE resources in one way or another. In 2015, towards the end of the project, the XSEDE User portal counted more than 20,000 active portal users, about half of which had allocations [16]. Many of the active users that didn't have allocations were registered with the portal in order to take part in training, workshops, or other XSEDE sponsored activities. There are users of XSEDE resources from every NSF directorate, including

Arts and Humanities, and the XDMoD metrics portal reports 147 fields of science charging resource use to the allocation reporting system over the past five years of the project.

That being said, the bulk of usage falls into a set of traditionally computational fields of study: across the five-year period of the XSEDE award, the top consumers of resources per job submitted were fields of science typically associated with the computational sciences. These were Engineering Systems, Galactic Astronomy, Polar Glaciology, Gravitational Physics, Global Atmospheric Research, Nuclear Physics, Polar Earth Sciences, Physical Oceanography, and Seismology. The exception to this was Law and Social Sciences, using the 5th most resources per job during this time. A chart of the largest per-job usage over the XSEDE period is displayed in Figure 5.2. Usage in these fields of science is overwhelmingly on the part of a single researcher in each, or sometimes completely the result of two researchers in that field. The outlier field of Law and Social Sciences results from the usage of XSEDE by Dov Cohen, a psychologist at UIUC.

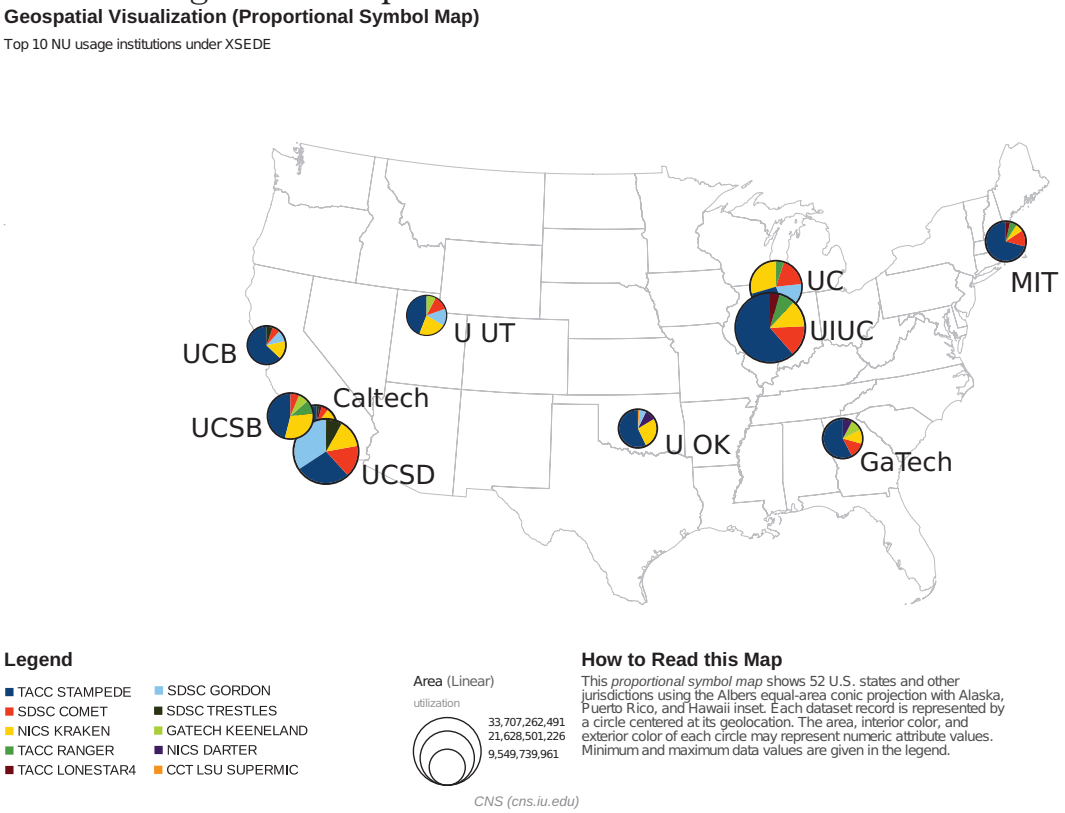
Looking at the largest institutions making use of XSEDE, it seems that proximity to XSEDE partner sites makes a difference, at least at the very top end of the scale. A map of the top institutional users of XSEDE is visible in Figure 5.3. The first three largest institutional users of XSEDE are at UIUC, UC, and UCSD, respectively. It would seem that having XSEDE staff nearby facilitates more usage and greater usage of XSEDE resources.

Figure 5.2: Field of Science usage during XSEDE



Despite the fact that neither UIUC nor UC have resources which feature into this group, the number of outreach activities and the draw of having a computational center pulls highly computational researchers to these institutions, and the ease of providing training and outreach activities locally also helps recruit users who might otherwise pass XSEDE by. It is important to note that the TACC Stampede resource makes up the greatest supplier of resources among the institutions in Figure 5.3, followed by NICS Kraken, and neither of these make the top 10 utilization list (TACC is in 18th place and NICS in 48th place), although users at SDSC show a clear preference for local resources: SDSC's systems make up more than two-thirds of the resources consumed at UCSD. Other XSEDE partners that show up within the top consumers of XSEDE NU's are Cornell (#12) and CMU (#17). Georgia Tech is another Service Provider making use of

Figure 5.3: Top 10 institutional users of XSEDE



local resources, as the Keeneland system provides significant resources.

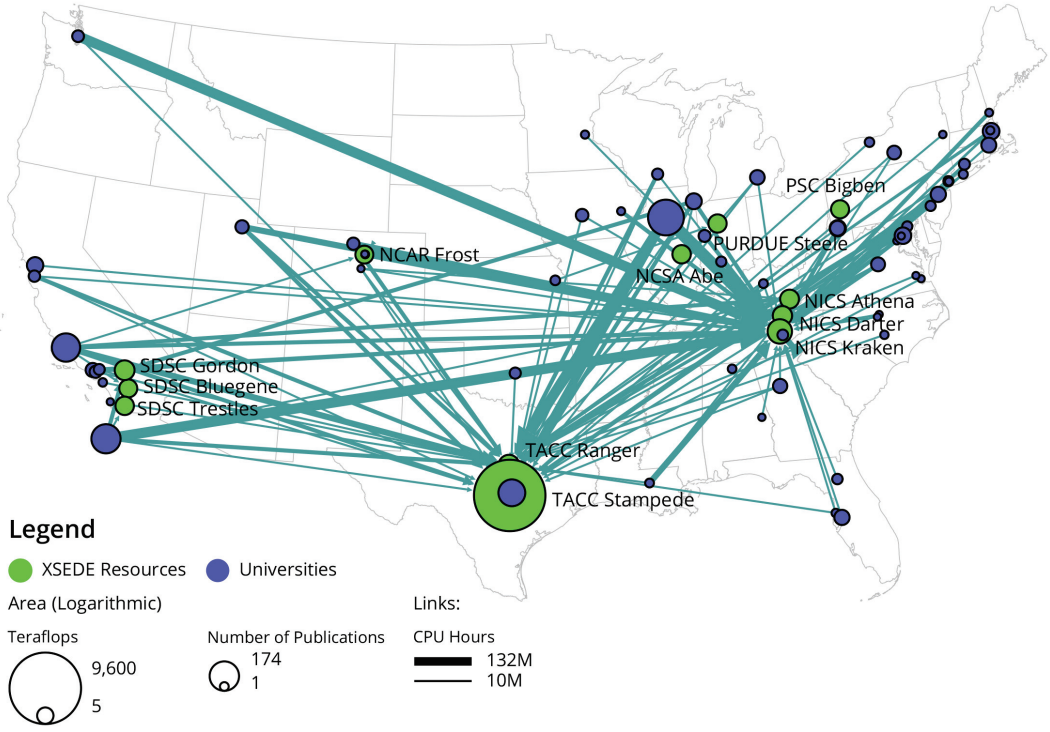
Not only is the top end of XSEDE utiliztion concentrated at sites with significant resources of their own, XSEDE utilization at the top end of the scale is clearly concentrated in the R1 universities of the United States. These institutions constitute a large percentage of the most active and largest users of XSEDE resources. Figure 5.4 shows the location of projects using more than 10M normalized CPU hours and the resources utilized by these institutions. When I presented this map to a member of the XSEDE allocations team, he remarked on how concentrated usage is among R1

sites and how little there is from states without significant resources at higher education institutions. Relatively few institutions represented in the figure come from states which have been recognized as beneficiaries of the NSF's Experimental Program to Stimulate Competitive Research (EPSCoR), which invests additional funds in support of research activities in order to enhance the research capabilities and provide significant avenues for scholarly advancement in these states [2]. Alabama, Colorado, Iowa, Kansas, Louisiana, Oklahoma, and Tennesee are prominent EPSCoR states with usage on the map, but this is less than one in five EPSCoR states making significant use of XSEDE resources. While, as I describe below, individual scientists with one or two projects in a field of science can drive some of the top usage of XSEDE, it does not appear that states under the EPSCoR program generate usage on part with the other users of XSEDE resources.

5.2.2 The "Long Tail" of XSEDE

Discussing the change in usage of XSEDE with my informants, it is clear that XSEDE staff feel that a change is in progress. There is signficant discussion within XSEDE about the development of science gateway usage, which became the most frequent means of submitting jobs within the XSEDE project, about new scientific software being turned at large data sets, notably bioinformatics software, which is not built to take advantage of HPC resources, and means of dealing with data sets in structured for-

Figure 5.4: Map of XSEDE utilization and resources)

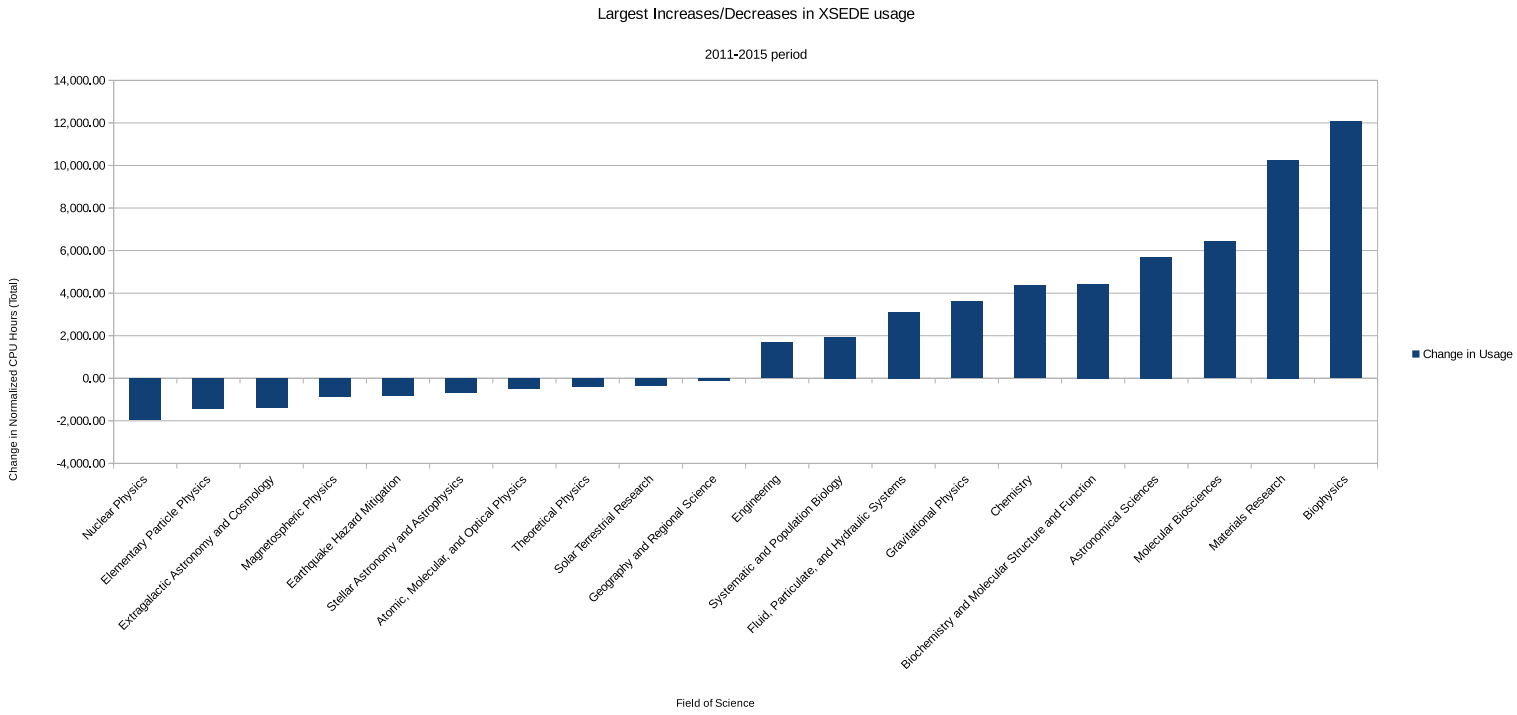


mats such as transactional databases, or clustered storage systems such as Apache Hadoop or NoSQL stores. As Hallam Stevens notes [146], bioinformatics research altered its questions in order to take advantage of computational resources, but many of the codes developed under bioinformatics' transition to computing were written for personal computers based on programming languages such as python and R that were not originally created to work on extremely large database sets. A frequent issue heard when discussing supporting bioinformatics research is the researcher or graduate student who writes a code that tries to load entire terabytes of data into memory, with the natural assumption that distributed systems with terabytes of memory should support what the coder's workstation cannot, and a difficult adaptation to the XSEDE environment. The proceedings of the TeraGrid 2011 Conference and successive XSEDE annual conferences show particular concern with adapting to new codes in this way. While these proceedings include science track papers on the large scale use of distributed resources that was typical of TeraGrid and continues to be a large portion of today's XSEDE utilization, there is also significant space dedicated towards questions about providing services for these types of codes, how to adapt them to XSEDE SP systems, and how to facilitate their use of resources. These new disciplines and techniques have earned the name the "long tail" of computational sciences, in that they may not be large-scale users, but still provide significant discoveries across the spectrum of usage, as Geoffrey Fox notes in his study of

science across the Branscomb Pyramid [10].

One way to look at changing fields of science across the XSEDE project is to examine usage changes by disciplines. Figure 5.5 shows the 10 fields of science with the largest drop and the 10 fields of science with the largest increase in normalized units of usage across the timeframe of the XSEDE grant. All of the disciplines with less utilization are traditional highly computational disciplines, such as Nuclear Physics, Particle Physics, and Astronomy. Most of the large gains are also from standard computational disciplines, although the largest gain overall is Biophysics and Molecular Biosciences, both of which are fairly new to large-scale computational workloads. This confirms the findings of Furlani et al, which identified similar trends up to 2013 [68]. Other fields which increased during the XSEDE project were Biochemistry and Molecular Structure and Function and Systematic and Population Biology. With four out of the ten largest gains in usage going to biological sciences, it appears that bioinformatics research has effectively made the switch to large scale computational infrastructure. It may be the case that the time to adapt to XSEDE resources is sufficient that new disciplines have a significant amount of lag before adoption can be complete. In 2011 at the beginning of XSEDE, there was considerable discussion about the rise of bioinformatics research as computational consumers, and these changes would seem to indicate that the rise has indeed occurred, over a five-year stretch. What remains to be seen is the appearance of other “long-tail” disciplines which

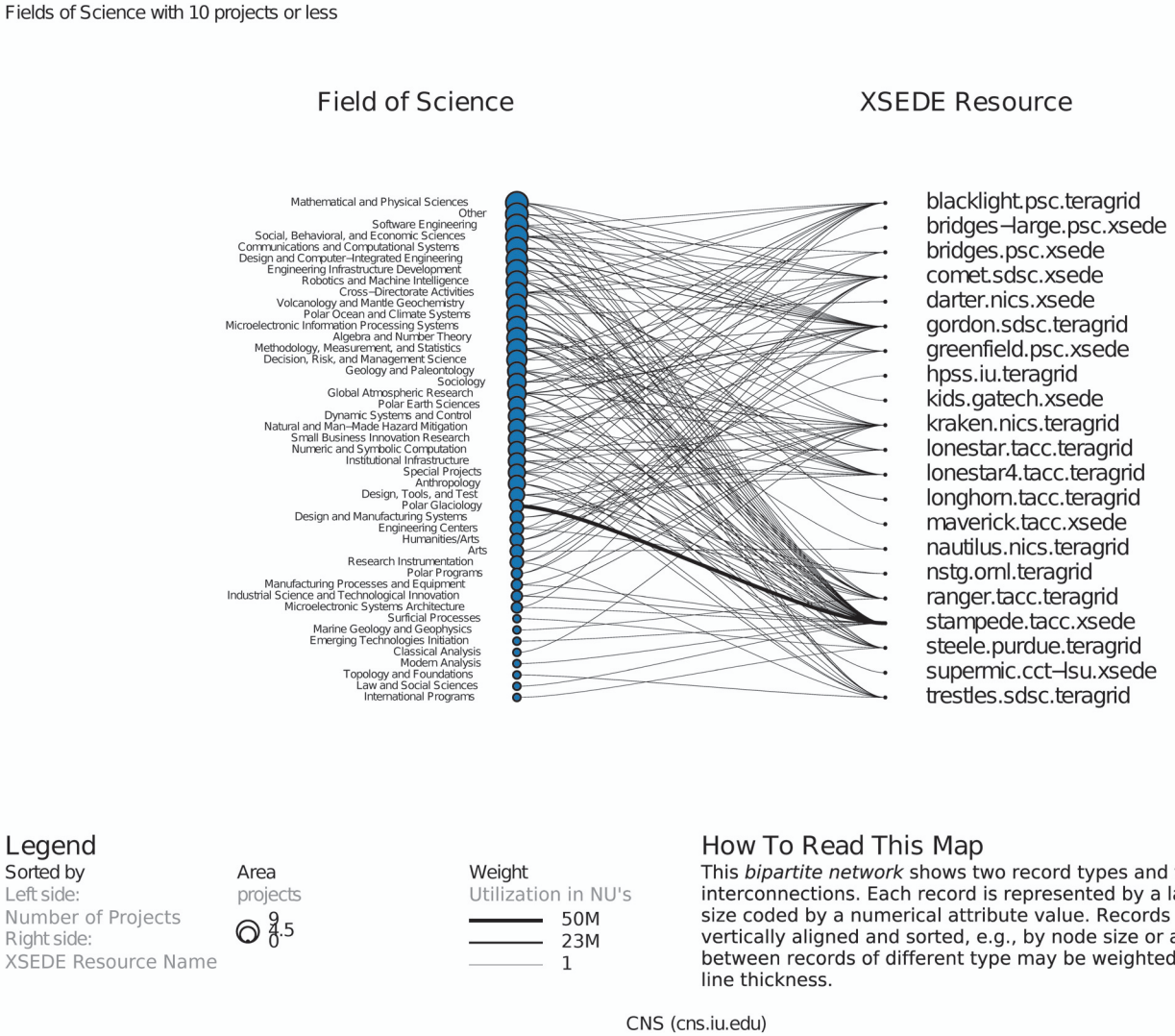
Figure 5.5: Changes in utilization by field of science, 2011-2016



may be arising over time.

This “long tail” of science encompasses a large number of activities. One way to classify it is simply to identify those fields which are less active than traditional fields. It may be possible to identify the members of this “long tail” by looking at fields of science with the fewest number of XSEDE projects. The range and breadth of XSEDE projects can make examining XSEDE usage patterns difficult, but it is possible to limit the elements involved. Figure 5.6 shows the NSF fields of science with fewer than 10 XSEDE projects, meaning that few Principal Investigators have in these fields have engaged with XSEDE. These projects range across many different disciplines, but notably include examples from the social sciences and the humanities. Some of the disciplines with few projects are also the disciplines with a considerable amount of usage, most notably Polar Glaciology shows up in the list of fields with fewest projects, as well as the top per-job fields. These fields with few PI’s working with XSEDE are still deriving significant usage from the resources, and can even drive the top utilization of XSEDE.

Figure 5.6: Looking for the long tail in XSEDE projects



In terms of the "long tail" as it represents new types of analysis and new disciplines, the usage seems to indicate a lack of these kinds of specialized computational analyses. While XSEDE SPs are providing resources for new fields (note the prominence of the "Other" field in Figure 5.6), the bulk of XSEDE service delivery has been for its traditional audience. Examining the XSEDE software search tool on the software search portal, there are a large number of bioinformatics codes available, as well as libraries such as hdf (a big-data file format) and big-data-r (statistics packages for handling big data), although usage of libraries on XSEDE is difficult to track. Bioinformatics software clearly has significant uptake on XSEDE, based on the changes in usage over the 5-year XSEDE period. However, these new ways of storing and accessing data appear to be slower in taking hold. It may be the case that some of these users are simply identifying other resources for carrying out analyses. One of my respondents, who is a former researcher and now academic administrator in charge of identifying computational resources for an institution with a considerable bioinformatics and big-data research agenda, stated simply, that "the long tail has already left", meaning that, rather than adapt their forms of computing to XSEDE and work out how to do batch job submission and similar activities, these researchers implement via cloud-based technologies such as Amazon EC2 or Microsoft Azure. For these researchers, it appears to be easier to pay for compute time through a third-party provider and run on systems that are more familiar than it is

to create an XSEDE startup allocation, adapt codes to work on XSEDE systems (sometimes with considerable data transfer requirements), and submit jobs to batch systems and wait for their completion. Another informant bore this out, when he told me that he was strongly considering leaving his position as a faculty member in order to pursue research in predictive analytics and statistical modeling. While the informant had completed what he regarded as good work on XSEDE resources thanks to a student fellowship, he regarded the new areas of data analytics to be more exciting and more renumerative than his current position as a faculty member. Not only may the long tail be outside of XSEDE, it may be the case that it is outside of academia completely.

5.2.3 How users leverage XSEDE

Further discussion within the cyberinfrastructure community and within cyberinfrastructure outreach organizations such as SURA frequently centers on how to provide services that are usable by universities that have less resources than the typical researcher on XSEDE. Part of XSEDE's mission to create a national research system is intended to even out distributional issues with open cyberinfrastructure. If a researcher needs to make use of advanced resources which are not available at their own institution, they can avail themselves of XSEDE resources. The XSEDE staff I met with who are responsible for broadening participation and bringing new communities of users to the project had few illusions about the

readiness of their audiences for engaging with XSEDE. Most of the activities with these communities revolve around providing a basis in computational techniques that can be carried out on personal laptops or lab computers. XSEDE resources are utilized to provide training, but the importance of the experience is to provide a computational basis with which these researchers can engage with their field of study currently, positioning XSEDE as a potential future resource.

My conversations with XSEDE users bore this out as well. I spoke to a number of users and campus champions who were appreciative of and extremely positive about XSEDE's offerings, but who admitted that their current set of research projects didn't make use of XSEDE allocations. However, these users were extremely savvy about the benefits of participating with XSEDE activities. One user mentioned her participation with XSEDE and plans to use SP resources in her proposal to the NSF for another research activity, which was subsequently funded. Another user made use of XSEDE training materials and supplementary activities, including an award for student funding, to further his research agenda, and did transition to carrying out research in XSEDE. Another respondent, a campus cyberinfrastructure professional, discussed working together on a pilot project in order to get firewall changes made at his institution which would be beneficial to the efforts of the researchers he serves.

Respondents made clear in interviews that while cyberinfrastructure was vital to their research aims, that their relationship with XSEDE was

not a collaborative one – that is, XSEDE does not appear to inform the science that it supports. Rather, the project provides a set of resources, of various kinds, not just computational, that researchers make use of to support their work. No respondent had noted that the nature of their inquiries might have changed based on what types of analysis and what tools were available to them in XSEDE.

All of the preceeding denotes a user base that is firmly rooted in instrumental use of XSEDE. Whether that is as a research tool, or as a tool for leverage to affect other goals in their research agenda, depends largely on the individual researcher's situation. Users who from traditionally computational fields, who are established in their usage patterns, have adapted their tactics to include not just the use of XSEDE resources, but they have built their credibility with the NSF, with other researchers, and with individuals at their own institutions. In the case of the researcher who used her experiences working with XSEDE's training programs, she viewed taking advantage of an existing NSF program to improve her computational skills and displaying a willingness to make use of XSEDE resources to support other NSF-funded research as giving her the credibility and the legitimacy to successfully propose for further NSF resources. The use of national computational resources is strategic for the researcher proposing for a grant. Rather than proposing to purchase her own computers and manage them, she can cite her other activities funded by the NSF for the purpose of supporting her research, and spend more time on the research

activity. Like many of the researchers at this scale of engagement, interacting with XSEDE and a declaration of intent to use XSEDE resources when the time is right is a means that researchers have to build their Latourian credibility with the NSF.

Respondents in XSEDE staff who are responsible for engagement activities are familiar with this situation. In speaking with staff charged with providing broader engagement activities, I learned that the bulk of the broader community that NSF mandates XSEDE support is not prepared to engage with the XSEDE infrastructure, either due to lack of resources for training at local institutions or lack of sufficient technical skill training in graduate programs. Despite this, through training and informational programs, XSEDE can support these researchers engaging in their own work. Working with XSEDE programs and participating in XSEDE initiatives lends legitimacy to these users' activities, no matter what activities or on what scale they engage with the project.

5.2.4 Researcher tactics

Researchers making use of XSEDE in traditional ways also noted that they had tactics for dealing with the XSEDE organization to get the most value possible out of the organization. One respondent, who is a developer of a molecular dynamics package run on a large number of XSEDE resources and who uses this software on a these same resources, described to me the development of tactics for dealing with the XSEDE Allocations pro-

cess. XSEDE research allocations requests are peer-reviewed for merit by a group of volunteers who meet on a quarterly basis. The XSEDE Research Allocations Committee (XRAC) awards those requests with sufficient scientific merit, as well as appropriateness of analyses to resources requested, and the feasibility of the request to return scientific results. As part of my observations I attended a meeting of the XRAC and was fascinated by the debate over awarding allocations. The XRAC, after a meeting including updates about newly-available resources and information about the number of available NU's to be awarded in the current round, reviews each of the allocations request in turn, and decides whether or not to award. At the end of that process, the awardees are allotted amounts of the award. Typically requests outnumber available NU's for allocation by about 3:2, and the typical solution is to reduce awarded allocations accordingly, depending on the decisions of the Allocations Committee. Given the demand for resources, and the fact that many researchers writing allocations requests are necessarily skilled in describing their work in the best possible terms, the XRAC tends to look for reasons that might disqualify an allocation. On the decision about whether to award, generally an XRAC member from the same field as the proposing researcher will present their opinion about the request. The members of the XRAC meeting I attended discussed merit and feasibility of requests, but the conversations also tended to range about more logistical concerns as well. XRAC members also reviewed the proposer's publishing history, whether

they had cited XSEDE as supporting previous work, whether or not it was the PI or graduate students were completing the work, or if an allocation request was legitimately in need of the amount requested. Sometimes similar requests were called out for being multiple members of the same lab or project requesting additional resources in order to get around policy limits.

My respondent the molecular dynamics developer confirmed that some of the requesting researchers certainly knew the tactics of the allocations committee and employed their own tactics to get allocations approved with as little reduction in resources aas possible. While having multiple members of the same research group request allocations was regarded as too easy for the XRAC to identify, generally inflating the amount of resources requested, with the rationale that all requests get reduced by some amount, was cited as one way to get the desired results. Another tactic was to omit certain details from the application. My respondent noted that requests that described NIH as well as NSF support for computation were frequently turned down, with the reasoning that researchers could get their computational resources from NIH rather than making use of overcommitted XSEDE resources. The picture that emerged from talking with my respondent and watching the XRAC make its determination was of two sides, largely communicating via the proposal and review documents, making use of limited information in order to make determinations about resource allocations. Researchers could build credibility

with the XRAC, most frequently by demonstrating prior work on XSEDE, or else lacking that, making use of an XSEDE Startup Allocation in order to demonstrate running and scaling of codes, by referencing XSEDE in their publications as providing resources to support their work. Notable requests that I saw disapproved during the XRAC meeting seemed to be those that reduced credibility of the requesting researchers, either by engaging in clear behaviors to receive more resources, or by publishing too slowly for the XRAC's standards (the XRAC viewed this as "wasting resources" without producing results, although no standard for producing results is part of the allocations policy or broader scientific work). Certainly attempts to game the peer-review system in the process of publishing in scientific journals is nothing new, tactics for getting journal articles accepted range from relatively innocuous to the seriously unethical [58]. While it is extremely doubtful that collusion or reviewing rings exist within the XRAC – the membership of the committee is much more stable than that of a journal, and the members conduct reviews all at once in the same room – it also appears that there are ways for users to get better access to resources, based on the way that they present their other support (such as omitting NIH grants) and asking for more than they expect to receive.

Another reason to work with XSEDE frequently cited by users of the organization is the ability to network and gain experience and information from other XSEDE users as well as staff members. The most prominent example of this is the Campus Champions group, whose member-

ship ranges from the cyberinfrastructure administrators at some universities to faculty or even non-technical administrative staff. Some are users of XSEDE resources, some are charged with helping faculty get a start on making use of resources, and some combine a number of roles together. This group carries out a lively ongoing discussion via email list and in-person workshops and at the annual XSEDE meeting. The topics range from hardware acquisition and implementation to application tuning, user support and facilitation. The Champions are the most intensely-connected group of XSEDE users, but other XSEDE users I approached noted the opportunities for interaction with other XSEDE users. What did seem prominent was that users felt singular in their use of XSEDE in regards to other users at their own institution or in their department. Most of the respondents I discussed XSEDE usage with might work with other faculty on the same XSEDE project, but few had much to say about getting other faculty members to make the transition. One respondent outright said that the other physics faculty at his institution had no interest in learning how to use resources and to do more computational investigations. It seems that XSEDE usership, or perhaps engaging in computational approaches more generally, represents a particular choice for researchers that determines some of their course of inquiry, and not all elect to make use of resources simply because they are on offer for free. For the researchers that do engage with XSEDE, however, there appears to be ample opportunity for collaboration and for networking which

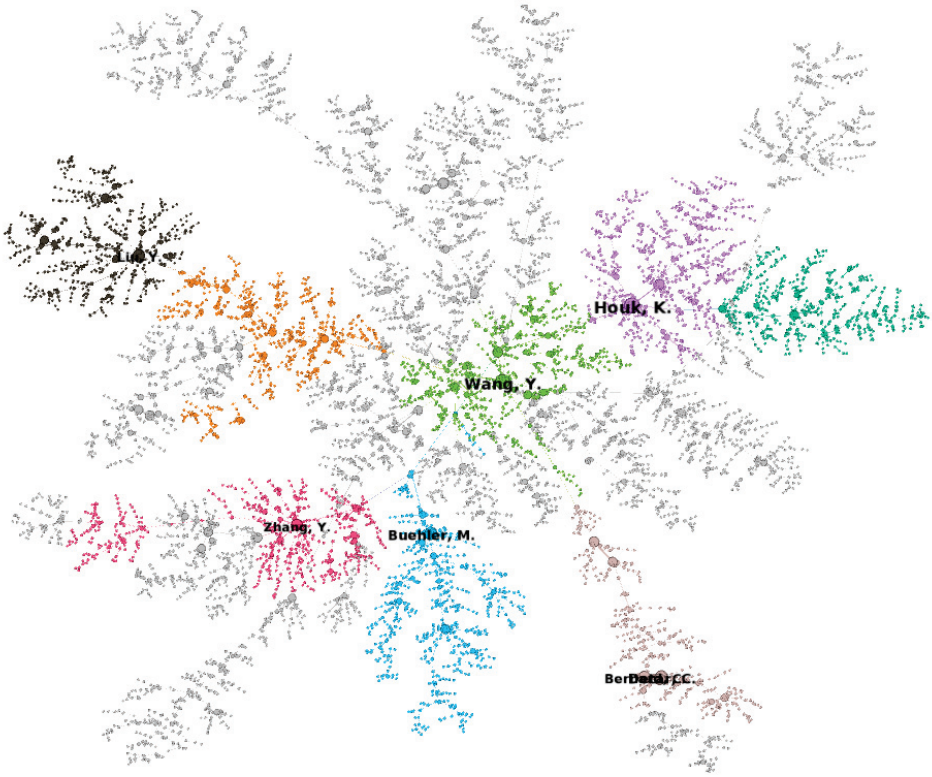
results in further research opportunities.

Turning a to a quantitative examination of relationships between XSEDE users, an examination of XSEDE outputs in the form of publications is informative. One way to quantitatively represent relationships between researchers is to make use of coauthorship networks in which researchers who are authors on publications with each other are linked by their common publications. The network of XSEDE researchers who publish work based on analyses carried out on XSEDE resources is extensive. The largest fully-connected subcomponent of the XSEDE coauthorship network, based on all researchers who provided publication information to the project, is shown in Figure 5.7, which is the result of my analysis of the publication data. This network has 9,256 authors, with each author in the network associating on average with two other authors (average degree of the network is 2, average weighted degree is 3.715). Modularity clustering in the network that is larger than 5% is shown by color, with the XSEDE authors with the largest number of published works of the eight communities detected of this size shown. Modularity clustering identifies these 8 communities as having more common coauthorship interactions with each other than with those outside of their communities.

5.3 Understanding XSEDE and the CI community

As described under Section 5.1.3, the beginnings of XSEDE were rooted in adaptation. Firstly, the XSEDE project as proposed was modified to

Figure 5.7: The XSEDE Coauthorship Network



incorporate elements of the XROADS proposal. Shortly after its inception, the project embarked upon a set of activities to provide additional governance and monitoring functions which would provide stakeholders additional information about XSEDE's execution of its activities. The architectural process adopted by XSEDE with input from the Software Engineering Institute were focused on the provision of capabilities that were rooted in user requirements. Over the course of executing the project, the evaluation team has conducted annual satisfaction surveys, staff climate surveys, and surveys about individual elements of XSEDE for incorporation in the XSEDE management structure. XSEDE has been a project that has taken external scrutiny to heart and made numerous changes to adapt.

Over the course of the project, emphasis has moved from the development and "hardening", or increasing the robustness and security, of software for distributed computing to responsiveness to user needs to effectively describing XSEDE's return on investment as compared to returning to the days of the Supercomputing Centers Program, when no central coordination between centers took place and competition for both researchers and staff was intense [143]. Grant-funded projects tend to be sensitive to stakeholder requests for changes and for information, and the XSEDE organization does not differ from the norm significantly. The Cyberinfrastructure community does frame itself under multiple "crises of the community" which may lend a sense of urgency to the project, and

the proposed changes.

5.3.1 Cyberinfrastructure Crises

As noted by Ensmenger, the computer industry has long framed its activities in terms of "crises" that it faces, initially in terms of a software crisis stemming from a lack of programmers [56]. The cyberinfrastructure community is not different in this respect. While attending XSEDE and NSF workshops, I heard about two crises that the cyberinfrastructure community framed for itself. Most often I attended presentations that started off with description of crises of capacity: the demand for computational capacity, and even more frequently, I participated in conversations about the ways to manage, share, and transfer data. Data storage capacity was a subject of discussion both within the XSEDE project at the first four quarterly meetings I attended in the form of discussions about an "XSEDE-wide file system". The pilot project I was involved with was also aimed at making data sharing between systems transparent. I also attended discussions at two CASC and a further NSF workshop that focused on questions of developing data management plans, supporting the need for better tools to handle large amounts of data and metadata. Meetings among cyberinfrastructure leadership that I attended described computing professionals facing the "data deluge" which would overwhelm the current means for storing and retrieving data. This is largely the rationale provided in the periodic reports on the need for providing HPC

resources to the research community [84, 96, 26, 35]. While the demand for resources continues to outstrip supply, this is not a situation that is expected to change, barring a radical reconfiguration of government expenditures.

A second crisis manifested in the inadequate workforce for cyberinfrastructure. There is a fair amount of agreement within the cyberinfrastructure community that there are insufficient professional staff and managers to meet existing needs, and that there is no training and education pipeline which adequately prepares students to engage with the profession [135, 114, 32]. At one of the meetings I attended for cyberinfrastructure professionals, the leader asked everyone in the room under 40 to raise their hand. In a room of about 50 attendees, there were fewer than five who did. The cyberinfrastructure workforce crisis is akin to the oft-mentioned workforce issues in Science, Technology, Engineering, and Math (STEM). The STEM workforce crisis is a matter of considerable debate around the extent of workforce needs and the policy initiatives which best address the situation [168, 133, 107]. The difference for the cyberinfrastructure workforce is that STEM degrees abound, while for the cyberinfrastructure professional there is no formal route to enter the field via a degree program or sub-program. Most of the people I met who work for XSEDE, the Open Science Grid, or at other projects, and most of my informants were either products of STEM education who "diverted" in the words of one, or computer science students who came to HPC instead

of going into industry or pure computer science work. Many of the of my co-workers and co-managers within XSEDE were trained in computational science disciplines, who noted that they felt they had an aptitude for the computational and began looking for opportunities within cyberinfrastructure rather than pursuing academic careers. Outside of a few courses on computational disciplines, and a few annual workshops on building skills for system administration, few venues exist for learning about performance tuning, advanced storage, or managing large scale federated systems. As such the XSEDE workforce has few formal structures by which to professionalize, enforce norms, and develop staff towards leadership. Certain activities, such as the annual SC conference, tend to be opportunities for staff to network, learn, and develop, but these opportunities are fairly few, and most of the XSEDE workforce needs to find its own way, with the stock of future leaders in fairly short supply. The risk facing the overall cyberinfrastructure community is that there will be serious shortfalls of leadership and difficulties creating and enforcing norms that provide common ground for the future leaders that eventually take the helms of the various supercomputing centers.

The environment created by these stories about crises is reinforced by the direction of NSF spending, in line with general science policy spending. During my first year working with XSEDE in 2011, I heard the phrase "flat is the new doubling", meaning that where science budgets had doubled in the past, the expectation was that researchers would be able to

accomplish the same or more with steady resources. Midway through the XSEDE project, that phrase was modified to "10% down is the new flat". Indeed, during my work with the project, I was requested to identify what activities would be cut based on a 10% funding reduction two times. The crises of resources and human resources within XSEDE, compounded with budget concerns, made for a particular urgency of the project requirements, and responsiveness to the NSF review process. Although there was never a threat that the project might be de-funded, everyone involved stressed the importance of performing well, in part to distinguish each of the participants for their future proposals and work with the NSF. Even XSEDE staff who showed signs of being more interested in their own center's fate were motivated to be responsive to requests, in order to build and retain credibility for their research organization. Whereas for XSEDE users, researchers leverage the organization in order to enhance individual credibility, the XSEDE partner organizations leveraged performance for stakeholders, visible service delivery, and science highlights in order to establish organizational credibility, preparing the partner organizations for the environment after XSEDE, as well as for proposed activities alongside XSEDE, such as Track 2 solicitations.

5.3.2 Adaptations in the Virtual Organization

These pressures on XSEDE resulted in organizational tendencies to adapt to the environmental requirements. XSEDE management spends consid-

erable time writing documents: quarterly and annual reviews, reponses to reviewer comments, annual program plans, and reponses to advisory board. Much of the work in order to become more responsive and manage the requests of the NSF focused on the improvement of processes. XSEDE constantly engaged in process improvement activities throughout the project. This section notes some of the activities adopted in order to improve the internal processes of XSEDE and responsiveness to XSEDE users as well as stakeholders.

First and foremost, XSEDE adopted a number of improvements in order to assist with collaborative work. The most basic of these dealt with an issue common in the TeraGrid as well as many virtual organzations: decisions made in distributed meeting settings were tentative. Until a central location for recording organizational decisions was made, leadership found themselves re-hashing the same conversations again and again. As members in a virtual organization, meeting settings were often conducted over the phone or by skype, reducing the immediacy of leadership interactions. Early on in the project, the need for documenting decision-making activities became clear. Documentation of processes and the activities in these processes made XSEDE able to pursue actions based on the decisions made by leadership and refer back to the documentation in the event of question or interpretation about those decisions. These ranged from relatively informal activities such as SMT meeting minutes kept by a program manager, to the documentation of software requests by the SD

& I team, which took place through a set of templated documents and issues in the Jira issue tracking system. These improvements allowed XSEDE to monitor and report on activities, and also allowed the organization to keep track of SP activities, in the case of slow or incorrect software implementations.

As mentioned in Section 5.1.3, XSEDE adopted a number of metrics and performance indicators, and through the course of the project, these metrics were adjusted on a regular basis. Frequently metrics were found to not count the output or outcome desired by the project, and needed to be revised. Similarly, targets for metrics often needed to be revised based on the performance of particular activities which were more successful than expected or encountered unanticipated obstacles. A portion of XSEDE management spent time on developing metrics for return on investment based on the cost of operations of individual centers versus the cost to centralize operations, as well as the value provided to users [150]. This exercise also surveyed center staff and users about their perceptions about XSEDE value, sometimes in telling ways. Not all of the center responses gave XSEDE central management positive valuation in terms of contributing to the overall research environment, indicating that for some, XSEDE was regarded as providing more drag than lift to efforts to build cyberinfrastructure. Nevertheless, in terms of cost models, XSEDE single operations center and security staff appear to provide significant savings over implementing similar functions across multiple centers.

XSEDE has also taken on the role of professionalizing both the cyberinfrastructure workforce as well as providing more education to researchers with normative outcomes, in the form of implementing badging and offering continuing education credits for engaging with XSEDE training. These activities contribute to the XSEDE workforce pipeline, in the case of badges, which are virtual signs of completion that can be used to show awareness or mastery of particular tools or skills. Continuing Education Credits are useful to researchers for demonstrating their own engagement with professional development by building similar skills. These efforts, developed by the TEOS group during the course of XSEDE, build both users and staff, and provide a way for those that have completed HPC training to have some sign of their learning activities.

XSEDE developments in terms of managing virtual organization: utilization of technical tools which support the collaborative (wiki, skype for business, atlassian tools).

Chapter 6

Conclusions

This research project attempts to make use of a singular amount of access to project resources, users, and staff members in order to get a view of the workings of a significant project funded by the NSF. XSEDE is intended to be a solution which provides resources available to all researchers regardless of their institution's means. Part of the difficulty the XSEDE project faces results from the fact that not all researchers are prepared to make use of XSEDE's offerings. Observations and interviews with both researchers and administrative staff identify that despite the lack of fit between some researchers and XSEDE's resources, they are able to make use of XSEDE in order to gain legitimacy in order to conduct other activities. Experienced users aware of XSEDE policy and activities maximize their access to resource by Users also make use of opportunities to collaborate with

Another way to think about crises of capacity is that there is a mismatch between the resources needed and those that are provided. If computational techniques provided the "third leg" of science in the form of

modeling and simulation, and XSEDE as an organization is firmly rooted in the provision of typical HPC resources, then the rise of big data techniques represents another type of inquiry. Big data investigation techniques are not yet well-fitted to the XSEDE modality, nor are the data storage and retrieval capabilities which support them, although it might be the case that a means of adapting these techniques to existing SP resources may arise. If XSEDE were to continue to provide resources as always, it would continue to support the numerical analysis community but have little to offer to the analytics community. XSEDE's is a large organization and NSF awards for Track 2 systems are a slow means of steering, but awards to research cloud systems rather than traditional HPC systems, associated awards for science gateways research and more sponsorship of the use of other computational capacities, including federated clouds, indicate that capability for the this new type of community is on the way.

6.1 Cyberinfrastructure trends

While there will always be a place for traditional HPC utilization and highly parallel systems, it appears that the trend of resources provided by XSEDE will continue to evolve. The shift in usage of XSEDE resources towards the biological fields of science and XSEDE's policies provide capabilities which answer user-driven requirements mean that the organization will be pushed to innovate and provide broader types of usage during

the next phase of the project. This is borne out by recent grants funded by the NSF. The NSF's last three awards for Track 2 systems has shown the NSF's preferences for a novel set of resources. The first of these is the Comet system at SDSC, a system which allows the creation of "virtual clusters" within a larger computational cluster. The PSC Bridges and Indiana University Jetstream system both provide research cloud capabilities, where users can start one or a number of virtual machine systems to carry out their analyses, then archive those systems once they are completed. The latest Track 1 award, the Stampede 2 system at TACC, provides 10 petaFLOPS of traditional parallel computing capability with accelerator-driven technologies. The first three systems provide flexibility and deliver a high degree of configurability to the user, while the last firmly supports high-powered parallel computing and the highly complex task of offloading algorithms to the secondary processing units. I have seen the discussion on the campus champion mailing list branch out into questions about the use of technologies such as containers which provide lightweight virtualization platforms.

Meanwhile, a broad range private providers, the most notable of which are Amazon EC2 and Microsoft Azure, continue to provide resources and support computation, providing cycles for hire. These systems are ultimately configurable and a broad range of solutions are implemented for automating the instantiation and running of these commercial viable systems. While these solutions do have an expense, they do not have the

allocations process that XSEDE resources have, and they have a considerable advantage in visibility over the XSEDE project. Large-scale use of commercial cloud platforms can also represent a significant expense.

The NSF supports a number of projects focused on the development of software to improve the user experience of cyberinfrastructure. Science gateways, which present a web-based front end to cyberinfrastructure resources, organized by discipline, provide access to a set of commonly-used codes, manage research data, and retrieve and share results. The NSF also funds initiatives to improve the reliability and standardization of software. While gateways provide a significant gain in ease of use, significantly reducing the amount of learning required to make use of resources, other initiatives to provide grid computing, such as the Unicore and Genesis II software projects, are still difficult to install, require special software to be installed on the resources, and present complex and obtuse user interfaces, hindering adoption. The main concern with NSF-funded software improvements is that, for the most part, these are not user-driven requirements, but proposed solutions, frequently from computer science researchers. The result is that there is a deep mismatch felt between most cyberinfrastructure software providers and users. One exception to this is the Globus project, which has adopted a web-based approach to data transfer that is relatively easy to parse and understand, and based on statistics incorporated into XSEDE reports grows by about 150 users every quarter. For the most part, however, cyberinfrastructure software

remains difficult to implement and use broadly.

6.2 XSEDE 2 and what comes after

In Chapter 7, I note that the TeraGrid was an effort to provide the basics of a distributed computing infrastructure, and XSEDE became a project around improving the service delivery processes of the infrastructure. Based on what I have seen since the end of the first iteration of XSEDE, which ended in June of 2016, there is reason to believe that cyberinfrastructure will continue along this path, refining activities and incorporating improvements from other NSF cyberinfrastructure software programs, until the next large-scale change in 2020.

In 2015, the NSF invited the leadership of XSEDE to propose a 5-year follow on to the XSEDE project. In order to maintain visibility and name recognition with its target audience, the project continues under the name of XSEDE (the formal title of the award is "XSEDE 2.0: Integrating, Enabling and Enhancing National Cyberinfrastructure). XSEDE 2.0 was shaped with a set of changes to the organization that would reduce the project's activities creating software and encourage the adoption of software funded by other NSF projects and elsewhere, reinforce the training and outreach activities, including broadening participation efforts, and continue to provide robust access to CI resources. Furthermore, XSEDE 2.0 has extended its integrative framework to allow more Service Providers, including those not funded by the NSF, to interoperate

with the project's resources. XSEDE 2.0 has also focused on providing frameworks that allow other resources to advertise capabilities available for use, even outside of XSEDE. In this, the project's focus has shifted from being the one provider for resources, to a participant in a larger fabric of resources.

Not long after XSEDE 2.0 was awarded, by August of 2016, when I attended the quarterly management meeting, conversations that I engaged in turned to what the NSF's plans would be for the next iteration of cyberinfrastructure organization. XSEDE is unusual among NSF awards for two reasons: its size in dollars and the role of the project. The oversight and management that XSEDE requires, considering that it is a virtual organization spread across 15 partner programs, is quite extensive for the foundation. Most of the leadership of XSEDE participating in these conversations in August 2016 opined that the NSF would most likely solicit proposals for 4 activities related to the cyberinfrastructure. These would probably end up being in the categories of operations, architecture, consulting and science gateways, and outreach. In addition, similar functions for other organizations, such as the Open Science Grid would be folded in, creating four linked organizations which would provide a "stack" of overall national cyberinfrastructure services, coordinated by the NSF. There are a number of concerns with this potential development.

If the NSF funds four separate grants to complete the activities outlined above, and incorporates other cyberinfrastructure projects in the

process, the burden will be on these organizations to coordinate with each other and to provide a consistent interface and environment throughout the transition of organizations. One of the first successes of the XSEDE project was the transition from TeraGrid to XSEDE operations without interruption in service for users. A national cyberinfrastructure stack to unite multiple projects would most likely take a considerable amount of critical work in order to continue operations without disturbing the activities of one or more of the original organizations. Furthermore, the operations organization would be forced to either support multiple differing ways to manage allocations of resources, or to find a way to coerce all of the member cyberinfrastructures to interoperate. The Service Provider model of XSEDE allows different centers to have different policies for their own systems but ensures a minimum level of interoperability, and perhaps this flexibility can be extended to further cyberinfrastructures. This would require those systems to become Service Providers to the future cyberinfrastructure stack, which may or may not be palatable to these projects. This is to say nothing of the cultural challenges of integrating multiple cyberinfrastructures with different norms about usage and technologies. Finally, the XSEDE organization has developed a significant understanding of processes and activities that support the conduct of science but also reporting that support to the NSF and other stakeholders. A divided organization will possibly fragment that collected knowledge among the four members of the cyberinfrastructure stack, or the institu-

tional knowledge to provide cyberinfrastructure services may dissipate in the new configuration, resulting in a lost investment for the NSF.

Based on the funding activities for XSEDE and other projects I have participated in, I have concerns that the Training and Outreach component in the new model could face difficulties based on the level of resources NSF allocates to it. An under-funded outreach organization will have problems recruiting users from the institutions which bring new perspectives and new disciplines, and this means that resource allocations will most likely remain centered over the states with traditionally strong research universities, rather than diversifying to those in EPSCoR states or to Minority Serving Institutions. If the aim of the NSF is to broaden participation in these types of cyberinfrastructure investments and create a pipeline of new and diverse users, careful allocation of resources needs to be made to this future outreach organization. The NSF can only truly affect national outcomes by ensuring that all have access to high-quality computational resources. Activities to broaden the XSEDE user base by introducing new communities and new disciplines must be genuine. One measure of this is the number of users who are introduced to the cyberinfrastructure have the opportunity to advance beyond their initial roles within the organization and are not restricted to being the "new users". It may be the case that the outreach program will have difficulties reaching across the cyberinfrastructure stack to operations and consulting organizations. In order to ensure that HPC training activities go smoothly, there

must be coordination between the trainers and those responsible for the system, and if future organizations introduce additional barriers to cooperation, training could suffer. A split but interdependent organizational format also poses challenges for the kinds of activities I identify between researchers and XSEDE. While it may be possible to leverage participation in training activities with other NSF solicitations, users may not be able to draw on their legitimacy with the outreach organization to effectively gain access to the other organizations in the future cyberinfrastructure stack.

6.3 Science Policy lessons

Other NSF grants which reach or surpass XSEDE's scale have been for investments which, like XSEDE, provide a number of services for researchers. These include large-scale centers, such as the National Center for Atmospheric Research and National Ecological Observatory Network, or instruments, such as the Large Synoptic Survey Telescope (LSST), or for Research Vessel operations, which are by their nature expensive undertakings. These initiatives, and many others being funded by the NSF at smaller levels and in different directorates, such as services to provide data, methods, and analyses to many researchers in the same field, represent a new direction for the NSF's activities. These types of projects are service delivery organizations for a broad set of investigations, rather than, as the centers, laboratories, and instruments listed above, the focal point for a particular discipline or area of inquiry. As in the progression

from TeraGrid to XSEDE to XSEDE 2.0, these projects are being redefined to become service delivery organizations. These future investments by the NSF must create the organizations that provide resources for a broad range of researchers across many disciplines, and they will be charged by the NSF to identify needs of these communities, capture metrics on the type and extend of activities engaged in, and they will need to provide extensive documentation on performance. While it may be arguable that the NSF, who is largely made up of highly-ranked scientists in their respective fields, collectively knows quite a bit about the conduct of good science, it remains to be seen that these same scientists have the same level of acumen concerning these types of service delivery organizations. It seems that the NSF will need to adapt and learn as it shifts the types of offerings it supports.

Furthermore, it appears that researchers are content to adapt private industry resources to their needs. More and more research-centered cloud-based activities are beginning to appear, and private offerings are contenders for ways to provide access to computational resources. While there are few offerings that provide actual access to HPC resources, new models are being offered that support activities with large data sets and the need for analytical processing of data. The "long tail" of science is incorporating these types of resources for its own usage, based on their flexibility, access, or simplicity of use. While next-generation HPC capabilities will remain the purview of the NSF (after that of the DoE and DoD

labs), there will be demand for other resources which meet the needs of research community for quick computational tasks.

6.4 Areas for further research

This research project has provided an in-depth perspective on the XSEDE project, supporting ethnographic description with quantitative data based on the use of XSEDE. There are a number of areas for future inquiry that may provide additional understanding of the role that these virtual organizations play in the NSF's ecosystem, and some direction about how best to create solicitations and structure organizations for the support of science. Based on the limited amount of change in XSEDE 2.0 from the initial organization, it may be the case that this organization is a less interesting subject for observation than its earlier incarnation. It may be more fruitful to turn a similar lens on the Open Science Grid, for example, to try and understand the forces that drive that project and the changes the OSG makes. That being said, the behavior of XSEDE 2.0 participants as the time for NSF to release its solicitation to replace XSEDE may be particularly informative in terms of what happens when the partners in such a large collaboration have increasing incentives to compete with each other rather than cooperate.

Firstly, understanding XSEDE as a virtual organization, capturing challenges in this form of work and the innovative responses that XSEDE has developed, gives considerable insight into the understanding of virtual or-

ganizations and what types of work cycle through the organization. The virtual organization alone is not the only level of inquiry that may provide useful information. A study of one of the centers that serves as an XSEDE level one Service Provider would be informative in grasping the attitudes of these centers towards XSEDE's consolidating function, as well as additional ways of interacting with stakeholders and managing users. In contrast to the central organization, there may be activities at the edge that are meaningful to understanding the relationship between resources and science. This type of investigation would inform the perspective of the virtual organization by providing more detail on what the participants and how they interact with XSEDE.

Secondly, development of the linkage between resources would be useful to inform science policy and other scientometric pursuits. Being able to draw relationships between the amount of resources utilized and the type and frequency of publication should provide NSF some guidance about the level and type of support. This also provides a possibility of monitoring the performance of a particular discipline based on resources consumed. Based on the last section about the introduction of new private resources, this also calls into question our understanding about what kind of researchers end up not making use of the resources as originally planned. In part this is an exercise in looking for the missing researchers, but careful surveying and conversations with researchers who might provide good examples of this kind of usage might help generate further ques-

tions about the utilization of private resources in comparison to the NSF's offerings, and to provide ideas about what barriers these researchers perceive in those offerings. There may be elements to the NSF-provided activities which are not immediately obvious from within the organization as they are to the researchers.

Bibliography

[1] Cyberinfrastructure: From Supercomputing to the TeraGrid | NSF - National Science Foundation.

[2] EPSCoR State Web Sites | NSF – National Science Foundation.

[3] f4transkript - faster transcription of interviews and recordings | audiotranscription.de.

[4] Facilities | U.S. DOE Office of Science (SC).

[5] High Performance Computing System Acquisition: Continuing the Building of a More Inclusive Computing Environment for Science and Engineering. | NSF - National Science Foundation.

[6] The Internet - The Launch of NSFNET.

[7] NSF Award Search: Advanced Search Results.

[8] nsf06573 Leadership-Class System Acquisition - Creating a Petascale Computing Environment for Science and Engineering | NSF - National Science Foundation.

[9] Programs | NSF - National Science Foundation.

[10] A Report of the National Science Foundation Advisory Committee for Cyberifnrastructure: Task Force on Cyberlearning and Workforce Development.

[11] Skype Call Recorder | Download.

[12] XSEDE | Overview.

[13] LHC Computing Grid Technical Design Report. Technical Report 1.0, June 2005.

[14] Cyberinfrastructure for 21st Century Science and Engineering, Advanced Computing Infrastructure: Vision and Strategic Plan. Technical report, National Science Foundation, 2012.

[15] A Vision and Strategy for Software for Science, Engineering, and Education: Cyberinfrastructure Framework for the 21st Century (NSF 12-113). Technical Report 12-113, National Science Foundation, 2012.

[16] XSEDE PY4 Annual Report. Technical report, July 2015.

[17] Peter A. Abrams. The Predictive Ability of Peer Review of Grant Proposals: The Case of Ecology and the US National Science Foundation. *Social Studies of Science*, 21(1):pp. 111–132, 1991.

[18] J. B. Adams. Megaloscience. *Science*, 148(3677):1560–1564, 1965.

[19] Pawan Agnihotri, Vijay K. Agarwala, Jeffrey J. Nucciarone, Kevin M. Morooney, and Chita Das. The Penn State computing condominium scheduling system. In *Proceedings of the 1998 ACM/IEEE conference on Supercomputing*, pages 1–23. IEEE Computer Society, 1998.

[20] Robert Agranoff and Michael McGuire. Encyclopedia of Public Administration and Public Policy. pages 552–557. Marcel Dekker, Inc, 2003.

[21] Saman Amarasinghe, Dan Campbell, William Carlson, Andrew Chien, William Dally, Elmootazbellah Elnohazy, Mary Hall, Robert Harrison, William Harrod, Kerry Hill, and others. Exascale software study: Software challenges in extreme scale systems. *DARPA IPTO, Air Force Research Labs, Tech. Rep*, 2009.

[22] D. P. Anderson. BOINC: a system for public-resource computing and storage. In *Fifth IEEE/ACM International Workshop on Grid Computing*, pages 4–10, November 2004.

[23] Amy Apon, Jeff Pummill, and Dana Brunson. Community Funding Models for Computational Resources. 2010.

[24] A. L. Beberg, D. L. Ensign, G. Jayachandran, S. Khaliq, and V. S. Pande. Folding@home: Lessons from eight years of volunteer distributed computing. In *2009 IEEE International Symposium on Parallel Distributed Processing*, pages 1–8, May 2009.

[25] Beth A. Bechky. Gaffers, gofers, and grips: Role-based coordination in temporary organizations. *Organization Science*, 17(1):3–21, 2006.

[26] Arden L. Bement Jr, Peter A. Kollman, Mary K. Vernon, John Hennessy, Andrew B. White Jr, John Ingram, Austin Schlumberger, William A. Wulf, Nathaniel Pitts, Robert Voigt, Edward F. Hayes, and Paul Young. Report of the Task Force on the Future of the NSF Supercomputer Centers Program. September 1995.

[27] Nicholas Berente, James Howison, John L King, Joel Cutcher-Gershenfeld, and Robert Pennington. Leading Cyberinfrastructure Enterprise: Value Propositions, Stakeholders, and Measurement. *Stakeholders, and Measurement (March 26, 2014)*, 2014.

[28] Jeremy M. Berg. Science policy: Well-funded investigators should receive extra scrutiny. *Nature*, 489(7415):203–203, September 2012.

[29] H. Russell Bernard and Clarence C. Gravlee. *Handbook of methods in cultural anthropology*. Rowman & Littlefield, 2014.

[30] David Berry. The computational turn: Thinking about the digital humanities. *Culture Machine*, 12, 2011.

[31] Frances S. Berry, Ralph S. Brower, Sang Ok Choi, Wendy Xinfang Goa, HeeSoun Jang, Myungjung Kwon, and Jessica Word. Three traditions of network research: What the public management research agenda can learn from other research communities. *Public administration review*, pages 539–552, 2004.

[32] Catherine Blake, Jeffrey M. Stanton, and AnnaLee Saxenian. Filling the workforce gap in data science and data analytics. 2013.

[33] James B. Bottum, Ruth Marinshaw, Henry Neeman, James Pepin, and J. Barr von Oehsen. The condo of condos. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, page 62. ACM, 2013.

[34] Geoffrey Bowker. Information mythology and infrastructure. *Information acumen: The understanding and use of knowledge in modern business*, pages 231–247, 1994.

[35] Lewis Branscomb, Theodore Belytschko, Peter Bridenbaugh, Teresa Chay, Jeff Dozier, Gary S. Grest, Edward F. Hayes, Barry Honig, Neal Lane, William Lester, Jr., Gregory J. McRae, James A. Sethian, Burton Smith, and Mary Vernon. *From desktop to teraflop: Exploiting the US lead in high performance computing*. National Science Foundation, 1993.

[36] Vannevar Bush. Science The Endless Frontier. Technical report, Office of Scientific Research and Development, 1945.

[37] Katy Börner. Science of Science Studies: Sci2 Tool. *Communications of the ACM*, 54(3):60–69, 2011.

[38] David F.J. Campbell. The Evaluation of University Research in the United Kingdom and the Netherlands, Germany and Austria. pages 98–131. Edward Elgar, 2003.

[39] James H. Capshew and Karen A. Rader. Big Science: Price to the Present. *Osiris*, 7:2–25, January 1992.

[40] C. Catlett, S. Goasguen, and J Cobb. TeraGrid Policy Management Framework, March 2006.

[41] C. et al Catlett. Teragrid: Analysis of organization, system architecture, and middleware enabling new types of applications. In *High Performance Computing (HPC) and Grids in Action*, number 16 in Advances in Parallel Computing. IOS Press, Amsterdam, 2008.

[42] Charles M. Vest, William J. Perry, Ray Kurzweil, and Calestous Juma. Engineering: Grand Challenges for the 21st Century | NSF - National Science Foundation, February 2008.

[43] Chen, Yu-Che and Knepper, Richard. Cyberinfrastructure for Collaborative Scientific Networks: Institutional Design and Management Strategies. In Chen, Yu-Che and Ahn, Michael, editors, *Routledge Handbook on Information Technology in Government*. Routledge.

[44] Ivan Chompalov, Joel Genuth, and Wesley Shrum. The organization of scientific collaborations. *Research Policy*, 31(5):749–767, 2002.

[45] Daryl Chubin and Edward Hackett. *Peerless Science: Peer Review and U.S. Science Policy*. State University of New York Press, 1990.

[46] Michael D. Cohen, James G. March, and Johan P. Olsen. A garbage can model of organizational choice. *Administrative science quarterly*, pages 1–25, 1972.

[47] Bozeman B. Brown E. Cozzens, S. Measuring and Ensuring Excellence in Government Laboratories: Practices in the United States. Technical report, Canadian Council of Science and Technology Advisors, 2001.

[48] Sharon M. Crook, Andrew P. Davison, and Hans E. Plesser. Learning from the Past: Approaches for Reproducibility in Computational Neuroscience. In James M. Bower, editor, *20 Years of Computational Neuroscience*, number 9 in Springer Series in Computational Neuroscience, pages 73–102. Springer New York, 2013. DOI: 10.1007/978-1-4614-1424-7_4.

[49] Maurice Crosland and Antonio Gálvez. The Emergence of Research Grants within the Prize System of the French Academy of Sciences, 1795-1914. *Social Studies of Science*, 19(1):71–100, 1989.

[50] Lorraine Daston and Peter Galison. The image of objectivity. *Representations*, 40:81–128, 1992.

[51] Andrew Davison. Automated Capture of Experiment Context for Easier Reproducibility in Computational Research. *Computing in Science & Engineering*, (4):48–56, July 2012.

[52] Derek John de Solla Price, Derek John de Solla Price, Derek John de Solla Price, and Derek John de Solla Price. *Little science, big science... and beyond*. Columbia University Press New York, 1986.

[53] Catello Di Martino, Zbigniew Kalbarczyk, Ravishankar K. Iyer, Fabio Baccanico, Joseph Fullop, and William Kramer. Lessons learned from the analysis of system failures at petascale: The case of blue waters. In *Dependable Systems and Networks (DSN), 2014 44th Annual IEEE/IFIP International Conference on*, pages 610–621. IEEE, 2014.

[54] P. N. Edwards. *A vast machine: Computer models, climate data, and the politics of global warming.* MIT Press, 2010.

[55] Paul N Edwards, Steven J Jackson, Geoffrey Bowker, and Cory Knobel. Understanding infrastructure: Dynamics, tensions, and design. 2007.

[56] Nathan L. Ensmenger. *The computer boys take over: Computers, programmers, and the politics of technical expertise.* Mit Press, 2012.

[57] Michael Feldman. IBM Bails on Blue Waters Supercomputer, August 2011.

[58] Cat Ferguson, Adam Marcus, and Ivan Oransky. Publishing: The peer-review scam. *Nature News*, 515(7528):480, November 2014.

[59] Thomas A. Finholt. Collaboratories. *Annual Review of Information Science and Technology*, 36(1):73–107, January 2002.

[60] Thomas A. Finholt and Jeremy P. Birnholtz. If we build it, will they come? The cultural challenges of cyberinfrastructure development. In *Managing nano-bio-info-cogno innovations*, pages 89–101. Springer, 2006.

[61] Thomas A. Finholt and Gary M. Olson. From laboratories to collaboratories: A new organizational form for scientific collaboration. *Psychological Science*, 8(1):28–36, 1997.

[62] Geraldine Fitzpatrick. Centres, peripheries and electronic communication: changing work practice boundaries. *Scandinavian Journal of Information Systems*, 12(1):6, 2000.

[63] Jean-Michel Fortin and David J. Currie. Big Science vs. Little Science: How Scientific Impact Scales with Funding. *PLOS ONE*, 8(6):e65263, June 2013.

[64] Ian Foster, Carl Kesselman, and Steven Tuecke. The anatomy of the grid: Enabling scalable virtual organizations. *International journal of high performance computing applications*, 15(3):200–222, 2001.

[65] Geoffrey C. Fox, Gregor von Laszewski, Javier Diaz, Kate Keahey, Jose Fortes, Renato Figueiredo, Shava Smallen, Warren Smith, and Andrew Grimshaw. Futuregrid: a reconfigurable testbed for cloud, hpc, and grid computing. In *Contemporary High Performance Computing: From Petascale toward Exascale*, pages 603–636. Chapman and Hall/CRC, 2013.

[66] H. George Frederickson. GOVERNANCE, GOVERNANCE EVERYWHERE. *The Oxford handbook of public management*, page 282, 2005.

[67] H. George Frederickson and Kevin B. Smith. Public Administration Theory Primer. *Boulder, CO: Westview*, 2003.

[68] Thomas R. Furlani, Barry L. Schneider, Matthew D. Jones, John Towns, David L. Hart, Steven M. Gallo, Robert L. DeLeon, Charng-Da Lu, Amin Ghadersohi, Ryan J. Gentner, Abani K. Patra, Gregor von Laszewski, Fugang Wang, Jeffrey T. Palmer, and Nikolay Simakov. Using XDMoD to Facilitate XSEDE Operations, Planning and Analysis. In *Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, XSEDE '13, pages 46:1–46:8, San Diego, California, USA, 2013. ACM.

[69] Peter Galison. *Image and logic: A material culture of microphysics*. University of Chicago Press, 1997.

[70] Peter Galison and Bruce William Hevly. *Big science: The growth of large-scale research*. Stanford University Press, 1992.

[71] Clara Garcia and Luis Sanz-Menendez. Competition for Funding as an Indicator of Research Competitiveness. *Scientometrics*, 64(3):271–300, 2005.

[72] Monica Gaughan and Barry Bozeman. Using curriculum vitae to compare some impacts of NSF research grants with research center funding. *Research Evaluation*, 11(1):17–26, 2002.

[73] Clifford Geertz. Thick description: Toward an interpretive theory of culture. *Readings in the philosophy of social science*, pages 213–231, 1994.

[74] Michael Gibbons and Björn Wittrock. Science as a commodity: Threats to the open community of scholars. 1985.

[75] Mark S. Granovetter. The strength of weak ties. *American journal of sociology*, 78(6):1360–1380, 1973.

[76] David H. Guston. Principal-agent theory and the structure of science policy, revisited:'Science in policy'and the US Report on Carcinogens. *Science and Public Policy*, 30(5):347–357, 2003.

[77] Warren O. Hagstrom. Competition in Science. *American Sociological Review*, 39(1):pp. 1–18, 1974.

[78] Robin Hanson. Patterns of Patronage: Why Grants Won Over Prizes in Science. *University of California, Berkeley*, page 11, 1998.

[79] Noriko Hara, Paul Solomon, Seung-Lye Kim, and Diane H. Sonnenwald. An emerging view of scientific collaboration: Scientists' perspectives on collaboration and factors that impact collaboration. *Journal of the American Society for Information science and Technology*, 54(10):952–965, 2003.

[80] David L Hart. Measuring TeraGrid: workload characterization for a high-performance computing federation. *International Journal of High Performance Computing Applications*, 25(4):451–465, 2011.

[81] Caroline Haythornthwaite, Karen J. Lunsford, Geoffrey C. Bowker, and Bertram C. Bruce. Challenges for research and practice in distributed, interdisciplinary collaboration. In *New infrastructures for knowledge production: Understanding e-science*, pages 143–166. 2006.

[82] Paul Hernandez. Baldrige Performance Excellence Program, December 2015.

[83] Bill Howe. Virtual Appliances, Cloud Computing, and Reproducible Research. *Computing in Science & Engineering*, (4):36–41, July 2012.

[84] Bill Joy, Ken Kennedy, Eric Benhamou, Vinton Cerf, Ching-Chih Chen, David Cooper, Steven Dorfman, David Dorman, Robert Ewalt,

David Farber, Sherrilynne Fuller, Hector Garcia-Molina, Susan Graham, James Gray, Daniel Hillis, Robert Kahn, John Miller, David Nagel, Raj Reddy, Edward Shortliffe, Larry Smarr, Joe Thompson, Leslie Vadasz, Andrew Viterbi, Steven Wallach, and Irving Wladawsky-Berger. Report to the President: Information Technology Research: Investing in Our Future. Technical report, President's Information Technology Advisory Committee, February 1999.

[85] Daniel S. Katz, Timothy G. Armstrong, Zhao Zhang, Michael Wilde, and Justin M. Wozniak. Many-task computing and Blue Waters. *arXiv preprint arXiv:1202.3943*, 2012.

[86] Kerk F. Kee, Lucy Cradduck, Bridget Blodgett, and Rami Olwan. Cyberinfrastructure inside out: definition and influences shaping its emergence, development, and implementation in the early 21st century. 2011.

[87] Erik-Hans Klijn and Chris Skelcher. Democracy and governance networks: compatible or not? *Public administration*, 85(3):587–608, 2007.

[88] Richard Knepper. The Shape of the TeraGrid: Analysis of TeraGrid Users and Projects As an Affiliation Network. In *Proceedings of the 2011 TeraGrid Conference: Extreme Digital Discovery*, TG '11, pages 54:1–54:6, Salt Lake City, Utah, 2011. ACM.

[89] Knepper, Richard. The XSEDE project: A Living Cyberinfrastructure. Leipzig, Germany, June 2014. Springer.

[90] Karin Knorr-Cetina. *The Manufacture of Knowledge: an Essay on the Constructivist and Contextual Nature of Science*. Pergamon Press, Oxford, 1981.

[91] Karin D. Knorr-Cetina. *The manufacture of knowledge: An essay on the constructivist and contextual nature of science*. Elsevier, 2013.

[92] Lew Kowarski. The Impact of Computers on Nuclear Science. In *Proceedings of the 1970 CERN Computing and Data Processing School*, pages 207–222, 1971.

[93] Lew Kowarski. New forms of organization in physical research after 1945. *Proceedings of the international school of physics "Enrico Fermi." Course LVU. History of twentieth century physics*, pages 370–90, 1977.

[94] Bruno Latour and Steve Woolgar. *Laboratory life: The construction of scientific facts*. Princeton University Press, 1986.

[95] Grit Laudel. The art of getting funded: how scientists adapt to their funding conditions. *Science and Public Policy*, 33(7):489–504, August 2006.

[96] Peter Lax, William Ballhaus, James C. Browne, Brice Carnahan, Michael Creutz, Richard Gallagher, Herbert Keller, John Killeen, Walter Macintyre, Steven Orszag, Werner Rheinboldt, Allan Robinson, Jacob Schwartz, Edward Wilson, and Kenneth Wilson. Report of the Panel on Large Scale Computing in Science and Engineering. Technical report, December 1982.

[97] Charlotte P. Lee, Paul Dourish, and Gloria Mark. The human infrastructure of cyberinfrastructure. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 483–492. ACM, 2006.

[98] K. LeRoux. *Service Contracting: A Local Government Guide.* International City/County Management Association, 2nd edition, 2007.

[99] Roland J. Liebert. Productivity, Favor, and Grants Among Scholars. *American Journal of Sociology*, 82(3):pp. 664–673, 1976.

[100] Milton Lomask. *A Minor Miracle: An Informal History of the National Science Foundation*. National Science Foundation, Washington, D.C., 1976.

[101] D.A. MacKenzie. *Knowing Machines: Essays on Technical Change*. MIT Press, 1998.

[102] Donald MacKenzie and Judy Wajcman. *The social shaping of technology*. Open university press, 1999.

[103] Joe Mambretti, Jim Chen, and Fei Yeh. Next generation clouds, the chameleon cloud testbed, and software defined networking (sdn). In *Cloud Computing Research and Innovation (ICCCRI), 2015 International Conference on*, pages 73–79. IEEE, 2015.

[104] Betty Matthews. ORNL debuts Titan supercomputer. Technical report.

[105] Amin Mazloumian, Dirk Helbing, Sergi Lozano, Robert P Light, and Katy Börner. Global multi-level analysis of the 'scientific food web'. *Scientific reports*, 3, 2013.

[106] James McClellan. *Science Reorganized: Scientific Socieities in the Eighteenth Century*. Columbia University Press, 1985.

[107] Heather Metcalf. Stuck in the pipeline: A critical review of STEM workforce literature. *InterActions: UCLA Journal of Education and Information Studies*, 6(2), 2010.

[108] H. Brinton Milward and Keith G. Provan. Managing networks effectively. In *National Public Management Research Conference, Georgetown University, Washington, DC October*, 2003.

[109] Ian M. Mitchell, Randall J. LeVeque, and Victoria Stodden. Reproducible research for scientific computing: Tools and strategies for changing the culture. *Computing in Science & Engineering*, (4):13–17, July 2012.

[110] Jacob Levy Moreno. Foundations of sociometry: An introduction. *Sociometry*, pages 15–35, 1941.

[111] Abbe Mowshowitz. Virtual organization: A vision of management in the information age. *The Information Society*, 10(4):267–288, October 1994.

[112] National Institutes of Health. Budget and Spending - NIH Research Portfolio Online Reporting Tools (RePORT), 2016.

[113] National Science Foundation. FY 2015 NSF Budget Request to Congress | NSF - National Science Foundation, 2015.

[114] Henry Neeman, Aaron Bergstrom, Dana Brunson, Carrie Ganote, Zane Gray, Brian Guilfoos, Robert Kalescky, Evan Lemley, Brian G. Moore, Sai Kumar Ramadugu, and others. The Advanced Cyberinfrastructure Research and Education Facilitators Virtual Residency: Toward a National Cyberinfrastructure Workforce. In *Proceedings of the XSEDE16 Conference on Diversity, Big Data, and Science at Scale*, page 57. ACM, 2016.

[115] Christena Nippert-Eng. *Watching Closely: A Guide to Ethnographic Observation*. Oxford University Press, 2015.

[116] NSTC. Assessing Fundamental Science: A report from the Subcommittee on Research, National Science and Technology Council Committee on Fundamental Science. Technical report, Washington, D.C., 1996.

[117] David E Nye. *American technological sublime*. MIT Press, 1996.

[118] Open Science Grid. OSG Connect - projects, 2017.

[119] Laurence J O'Toole and Kenneth J Meier. Desperately seeking Selznick: Cooptation and the dark side of public management in networks. *Public Administration Review*, 64(6):681–693, 2004.

[120] Laurence J. O'Toole Jr. Treating networks seriously: Practical and research-based agendas in public administration. *Public administration review*, pages 45–52, 1997.

[121] Dasgupta Partha and Paul A. David. Toward a new economics of science. *Research Policy*, 23(5):487 – 521, 1994.

[122] Roger D. Peng. Reproducible Research in Computational Science. *Science*, 334(6060):1226–1227, December 2011.

[123] Andrew Pickering. Big Science as a Form of Life. In Michelangelo De Maria, Mario Grilli, and Fabio Sebastiani, editors, *The Restructuring of Physical Sciences in Europe and the United States, 1945-1960*. World Scientific, 1989.

[124] Marlon E. Pierce, Suresh Marru, Lahiru Gunathilake, Don Kushan Wijeratne, Raminder Singh, Chathuri Wimalasena, Shameera Ratnayaka, and Sudhakar Pamidighantam. Apache Airavata: design

and directions of a science gateway framework. *Concurrency and Computation: Practice and Experience*, 27(16):4282–4291, 2015.

[125] Volkmar Pipek and Volker Wulf. Infrastructuring: Toward an integrated perspective on the design and use of information technology. *Journal of the Association for Information Systems*, 10(5):1, 2009.

[126] Ruth Pordes, Bill Kramer, Doug Olson, Miron Livny, Alain Roy, Paul Avery, Kent Blackburn, Torre Wenaus, Frank K. Wuerthwein, Rob Gardner, Mike Wilde, Alan Blatecky, John McGee, and Rob Quick. The Open Science Grid. *J.Phys.Conf.Ser.*, 78:012057, 2007.

[127] Don K Price. *The Scientific Estate*. Harvard College, 1965.

[128] Jerome R. Ravetz. Scientific knowledge and its social problems. 1971.

[129] D. Ribes and T.A. Finholt. The long now of technology infrastructure: articulating tensions in development. *Journal of the Association for Information Systems*, 10(5):375–398, 2009.

[130] Robert Ricci, Eric Eide, and CloudLab Team. Introducing CloudLab: Scientific infrastructure for advancing cloud architectures and applications. *; login:: the magazine of USENIX & SAGE*, 39(6):36–38, 2014.

[131] Horst WJ Rittel and Melvin M. Webber. Dilemmas in a general theory of planning. *Policy sciences*, 4(2):155–169, 1973.

[132] Lester M. Salamon, editor. *The Tools of Government: A Guide to the New Governance*. Oxford University Press, New York, 2002.

[133] Hal Salzman. What shortages? The real evidence about the STEM workforce. *Issues in Science and Technology*, 29(4):58–67, 2013.

[134] Stephen L. Schensul, Jean J. Schensul, and Margaret Diane LeCompte. *Essential ethnographic methods: Observations, interviews, and questionnaires*, volume 2. Rowman Altamira, 1999.

[135] J. M. Schopf. Sustainability and the Office of CyberInfrastructure. In *2009 Eighth IEEE International Symposium on Network Computing and Applications*, pages 1–3, July 2009.

[136] Science Gateways Community Institute. What is a Science Gateway: The Basics | ScienceGateways.org, 2017.

[137] Scientific Computing and Visualization Group and National Center for Supercomputing Applications. AMIE - Account Management Information Exchange.

[138] W. Richard Scott and Gerald Fredrick Davis. *Organizations and organizing: Rational, natural, and open system perspectives*. Prentice Hall, 2007.

[139] Steven Shapin, Simon Schaffer, and Thomas Hobbes. *Leviathan and the air-pump*. Br Soc Philosophy Sci, 1985.

[140] Philip Shapira and Stefan Kuhlman, editors. *Learning from Science and Technology Policy Evaluation*. Edward Elgar, Cheltenham, UK, 2003.

[141] Mark Sheddon, Ann Zimmerman, John Cobb, Dave Hart, Lex Lane, Scott Lathrop, Sergiu Sanielevici PSC, Kevin Walsh, and Dane Skow. Measuring TeraGrid Impact: Methods to Document Effects of TeraGrid Resources and Capabilities on Scientific Practice and Outcomes Impact Requirements Analysis Team.

[142] Shava Smallen, Catherine Olschanowsky, Kate Ericson, Pete Beckman, and Jennifer M. Schopf. The inca test harness and reporting framework. In *Supercomputing, 2004. Proceedings of the ACM/IEEE SC2004 Conference*, pages 55–55. IEEE, 2004.

[143] Larry Smarr. The Good, the Bad and the Ugly: Reflections on the NSF Supercomputer Center Program.

[144] S. Star and K. Ruhleder. Steps toward an ecology of infrastructure: design and access for large information spaces. *Information Systems Research*, 7(1):111, 1996.

[145] Susan Leigh Star. The ethnography of infrastructure. *American behavioral scientist*, 43(3):377–391, 1999.

[146] Hallam Stevens. *Life out of sequence: a data-driven history of bioinformatics*. University of Chicago Press, 2013.

[147] Craig A. Stewart, Timothy M. Cockerill, Ian Foster, David Hancock, Nirav Merchant, Edwin Skidmore, Daniel Stanzione, James Taylor, Steven Tuecke, George Turner, and others. Jetstream: a self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, page 29. ACM, 2015.

[148] Craig A. Stewart, Daniel S. Katz, David L. Hart, Dale Lantrip, D. Scott McCaulay, and Richard L. Moore. Technical Report: Survey of cyberinfrastructure needs and interests of NSF-funded principal investigators. January 2011.

[149] Craig A. Stewart, Stephen Simms, Beth Plale, Matthew Link, David Y. Hancock, and Geoffrey C. Fox. What is cyberinfrastructure. In *Proceedings of the 38th annual ACM SIGUCCS fall conference: navigation and discovery*, pages 37–44. ACM, 2010.

[150] Stewart, Craig, Roskies, Ralph, Knepper, Richard, Moore, Richard, Whitt, Justin, and Cockerill, Timothy. XSEDE Value Added, Cost Avoidance, and Return on Investment. *Proceedings of XSEDE 15 Conference*, July 2015.

[151] XSEDE Architecture Team. XSEDE Architecture Overview v2.0. September 2014.

[152] J. Torrellas. Architectures for Extreme-Scale Computing. *Computer*, 42(11):28–35, November 2009.

[153] John Towns. TeraGrid: Status and Challenges, April 2009.

[154] John Towns, Phil Andrews, Jay Boisseau, Ralph Roskies, Janet Brown, Kathlyn Boudwin, Tim Cockerill, Andrew Grimshaw, Patricia Kovatch, Scott Lathrop, Mike Levine, Sergiu Sanielevici, Dan Stanzione, and Kurt Wallnau. Project Summary: XSEDE: eXtreme Science and Engineering Discovery Environment. July 2010.

[155] Sharon Traweek. Big science and colonialist discourse: Building high-energy physics in Japan. *Big science: The growth of large-scale research*, pages 99–126, 1992.

[156] Merle A. Tuve. Is Science Too Big for the Scientists? *Saturday Review*, 6, 1959.

[157] University of Pittsburgh Office of Research. Federal Grant vs. Federal Contract Guidance, 2012.

[158] Moshe Y. Vardi. Science has only two legs. *Communications of the ACM*, 53(9):5–5, September 2010.

[159] John Walsh. NSF Peer Review Hearings: House Panel Starts with Critics. *Science*, 189(4201):pp. 435–437, 1975.

[160] Fugang Wang, Gregor von Laszewski, Geoffrey C. Fox, Thomas R. Furlani, Robert L. DeLeon, and Steven M. Gallo. Towards a Scientific Impact Measuring Framework for Large Computing Facilities - a Case Study on XSEDE. In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, XSEDE '14, pages 25:1–25:8, Atlanta, GA, USA, 2014. ACM.

[161] Stanley Wasserman and Katherine Faust. *Social network analysis: Methods and applications*, volume 8. Cambridge university press, 1994.

[162] Alvin M. Weinberg. Impact of large-scale science on the United States. *Science*, 134(3473):161–164, 1961.

[163] Alvin M. Weinberg. The federal laboratories and science education. *Science*, 136(3510):27–30, 1962.

[164] Norbert Wiener. Science: The megabuck era. *The New Republic*, 138:10–11, 1958.

[165] Jeannette M. Wing. Computational thinking. *Communications of the ACM*, 49(3):33–35, 2006.

[166] Langdon Winner. Upon opening the black box and finding it empty: Social constructivism and the philosophy of technology. *Science, Technology, & Human Values*, 18(3):362–378, 1993.

[167] XSEDE Project. XSEDE Revision Narrative. Technical report, XSEDE project, January 2011.

[168] Yi Xue and Richard C. Larson. STEM Crisis or STEM Surplus: Yes and Yes. *Monthly Lab. Rev.*, 138:1, 2015.

[169] Sierk Ybema, Dvora Yanow, Harry Wels, and Frans H. Kamsteeg. *Organizational ethnography: Studying the complexity of everyday life*. Sage, 2009.

[170] Charles N Yood. *Hybrid Zone: Computers and Science at Argonne National Laboratory 1946–1992*. Docent Press, 2013.

[171] Paul R. Zilsel. The Mass Production of Knowledge. *Bulletin of the Atomic Scientists*, 20(4):28–29, 1964.

[172] Ann Zimmerman and Thomas Finholt. Report from the TeraGrid Evaluation Study, Part 1: Project Findings. August 2008.

[173] Ann Zimmerman, Magia Krause, Katherine Lawrence, and Thomas Finholt. TeraGrid Evaluation Report, Part 2: Findings from the TeraGrid User Survey. July 2008.

Appendix A

Interview Questions

Starting questions for interviews with cyberinfrastructure users and staff

1. I'm going to ask you a few questions about your background in research and cyberinfrastructure, to understand what your experiences have been and what you have
   (a) Tell me a little about your background, where you went to school, and about your current job.
   (b) Tell me briefly how you're involved with national cyberinfrastructure (CI): XSEDE, Open Science Grid, or otherwise?
      - (if staff: What is your role within the project and what do you do?)
   (c) How did you become involved with computational infrastructure? With (CI) specifically?
   (d) How did you learn to start using (CI)?
   (e) Were there difficulties you encountered?
   (f) What did you do in order to overcome those difficulties?
      - (get help from others/read documentation/ support tickets/courses)
   (g) Was there anything about your own background that made it easier or harder to make use of (CI)?
   (h) Has your use of (CI) changed since you first engaged with it? How?
   (i) Has the use of (CI) by others in the community changed in the same time? How?
2. User Questions

(a) Do you feel that (CI) is important to your own work? How?

(b) Would more access to more (CI) resources that you currently use allow you to get more science done?

(c) Would different resources (cloud vs HTC vs HPC) allow you to get more science done?

(d) Are there other services that (CI) provides that give more benefits than just access to the resources?

(e) Do you feel like you are part of the general population of (CI) users? Do you feel like you belong to part of a group within the (CI) user base? Describe your group, what are its concerns and what are its values in relation to the project? Are you aware of other groups within the (CI) community?

(f) How do you perceive the allocation of resources within (CI)? Is it fair? Does your group (if identified) get an equitable share of the resources requested? Why? Do other groups receive a different share? Why?

(g) In the context of (CI) do you most often work alone, or with others, from the same group or from other groups, with staff members?

3. Staff Questions

(a) Does your background in (computational sciences, IT, other background from 1.a) support or affect your role in working on (CI)?

(b) Do you feel your support of (CI) contributes to the development of basic science? How?

(c) Does the current set of available resources meet the needs of the user community? Where does it not meet those needs?

(d) What role do you see (CI) playing in researcher activities – a partner, collaborator, instrument, or something else? Why? Should researchers be asked to cite or otherwise acknowledge (CI) in their publications?

(e) When discussing resource allocation, the high demand for resources compared to the available systems is often discussed. In your view, what makes a project more likely to receive an allocation (or other resources)? Are there factors which make the process difficult for new users?

(f) What do you think about new ways of accessing resources, such as via science gateways? What implications does this have for CI providers? For users?

4. Collaborative Science

   (a) Users:

      i. How would you characterize your own research work: largely your own work or conducted in collaboration with other researchers? In your broader field?
      ii. Do you use science gateways or other similar technologies (Prompt: such as HubZero or Cyverse) in your own research? Why or why not?
      iii. Do you understand science gateways as contributing to use of (CI)? How?
      iv. How do you understand analyses carried out in (CI) to be the product of collaborative or individual work?
      v. Do you feel that (CI) is a collaborative effort? How would you describe CI in relation to the research that you carry out? [prompt: as a tool? A partner? A co-creator? Something else?]

   (b) Staff:

      i. Do you see (CI) as enabling the collaborative practice of science? Why or why not?
      ii. In your work, do you engage with collaborative research projects? In what ways?
      iii. Are you involved in other collaborative efforts than (CI)?
      iv. How does your collaborative work align with competitive activities such as grant proposals and similar? Are there any particular actions that you take in order to manage the balance of these relationships?

5. Demographics of science questions

   (a) What are things that make it difficult to progress in your particular field of science?

   (b) What are things that make it difficult to make effective use of cyberinfrastructure?

(c) What do you see students struggle with in progressing within your field? With making use of cyberinfrastructure to get results in their research?

(d) Are there factors that make these barriers more difficult for some researchers than others? (prompt: resources of local institution, structural factors, fundamental education)

Richard Knepper

✉ *rknepper@iu.edu*
homes.soic.indiana.edu/rknepper
ORCID: 0000-0002-4296-9421

Professional Profile

- Senior Manager in Research IT with 15 years of experience consulting with and providing operations support for researchers at Indiana University and other institutions, including collaborating on grant projects
- Project leader in multiple NSF and NASA-funded projects, with roles including drafting proposals and statements of work, budgeting, managing operations staff, establishing and reporting metrics and key performance indicators. Works across multiple groups in Virtual Organizations to accomplish goals.
- Researcher in Cyberinfrastructure studies, examining the distribution of resources within high-performance computing environments supporting basic research

Areas of Expertise

Leadership	IT management and operations; Grant-funded project management; Project management and system implementation; Training and Development; Budgeting and Metrics
Technical	Linux/Unix System Administration; Apache/Tomcat; MySQL administration; Ticket tracking systems; Monitoring systems; Perl, Python, PHP
Analytics	Social Network Analysis; Univariate and Multivariate statistics; Linear Programming; R, Sci2, Gephi, Pajek

Professional Experience

2012-Present **Manager, Campus Bridging and Research Infrastructure**,
University Information Technology Services, Indiana University,
Co-Investigator for Indiana University role in XSEDE (Extreme Science and Engineering Discovery Environment, NSF OCI-1053575)
Co-Investigator for Indiana University EAGER: Best Practices and Models for Robust Cyberinfrastructure Software (NSF 1147606)
Principal Investigator for IU Subcontract to University of Kansas Center for the Remote Sensing of Ice Sheets/NASA: Operation Ice Bridge: A 3-year Airborne Mission for Earth's Polar Ice
Project management for research projects making use of IU's research cyberinfrastructure.

2007-2012 **Manager, Research Technologies Core Services**,
University Information Technology Services, Indiana University,
Systems analysis and integration for Polar Grids project (NSF CNS-0723054), including proposal budgeting, coordinating software and hardware implementation, planning and completion of fieldwork at Polar sites
Resource Provider Lead for FutureGrid project (NSF OCI-0910812), including chair of User Service Committee, management of support systems for FutureGrid users
Management of accounts management, accounting, and reporting for TeraGrid project, including compiling quarterly report information.

2005-2007 **Senior Unix Systems Analyst, Unix Systems Support Group**,
University Information Technology Services, Indiana University,
Support of faculty, students, and staff using Unix and Linux at IU
Curriculum planning and teaching of Unix System Administration Education Certification and Unix for Advanced Users classes .

2002-2005 **Lead Digital Media Analyst, Digital Media Network Services**,
University Information Technology Services, Indiana University,
System administration for VRVS reflector, EMC Clarion NAS device, streaming video servers
System administration for videoconferencing global management system, and associated services .

Education

2017 **PhD**, *School of Informatics and Computing*, Indiana University.
Social Informatics
Dissertation: Shifting modalities of use in the XSEDE project

1999-2002 **Master of Public Administration**, *School of Public and Environmental Affairs*, Indiana University.
Specialization: Information Systems

1999-2002 **Master of Arts, Polish Studies**, *Russian and East European Institute*, Indiana University.
Master's thesis: Evaluation of the ePolska Plan for Information Technology Development

1993-1997 **Bachelor of Science**, *New College of Florida*.

Selected Recent Publications

2016 Knepper, R. and Börner, K. (2016). **Comparing the Consumption of CPU Hours with Scientific Output for the Extreme Science and Engineering Discovery Environment (XSEDE)**. PLOS ONE *(in press)*
Knepper, R. and Chen, Yu-Che. (2016). **Situating Cyberinfrastructure in the Public Realm: The TeraGrid and XSEDE Projects.** *(accepted at the dg.o '16 conference, June 08-10, 2016)*

2015 Knepper, R., Standish, M., and Link, M. (2015). **Big data on ice: The forward observer system for in-flight synthetic aperture radar processing.** Procedia Computer Science, 51:1504-1513.
Stewart, C. A., Barnett, W. K., Wernert, E. A., Wernert, J. A., Welch, V., and Knepper, R. (2015). **Sustained software for cyberinfrastructure: Analyses of successful efforts with a focus on NSF-funded software.** In Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models, pages 63-72. ACM.

Stewart, C. A., Roskies, R., Knepper, R., Moore, R. L., Whitt, J., and Cockerill, T. M. (2015). **XSEDE value added, cost avoidance, and return on investment.** In Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure, page 23. ACM. *(Recipient of the Phil Andrews Best Technology Paper Award)*

2014 Fischer, J., Knepper, R., Standish, M., Stewart, C. A., Alvord, R., Lifka, D., Hallock, B., and Hazlewood, V. (2014). **Methods for creating xsede compatible clusters.** In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment, page 74. ACM.

Complete publications list available at http://homes.soic.indiana.edu/rknepper/publications