Experiences in network and data transfer across large virtual organizations—a retrospective

Kathy Benninger
benninge@psc.edu
Pittsburgh Supercomputing Center, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Chris Jordan
ctjordan@tacc.utexas.edu
Texas Advanced Computing Center
Austin, TX, USA

Michael Lambert
lambert@psc.edu
Pittsburgh Supercomputing Center, Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Lee Liming
lliming@uchicago.edu
University of Chicago
Chicago, Illinois, USA

Tabitha K. Samuel
tsamuel@utk.edu
National Institute for Computational Sciences, University of Tennessee, Knoxville
Oak Ridge, Tennessee, USA

David Wheeler
dwheeler@illinois.edu
National Center for Supercomputing Applications, University of Illinois
Urbana, Illinois, USA

ABSTRACT

The XSEDE Data Transfer Services (DTS) group focuses on streamlining and improving the data transfer experiences of the national academic research community, while also buttressing and future-proofing the underlying networks that support these transfers. In this paper, the DTS group shares how network and data transfer technologies have evolved over the past six years, with the backdrop of the Distributed Terascale Facility (DTF) and TeraGrid projects that served the national community before the advent of XSEDE. We delve into improvements, challenges, and trends in network and data transfer technologies, and the uses of these technologies in academic institutions across the country, which today translate into 100s of users of CI moving many terabytes each month. We also review the key lessons learned while serving the community in this regard, and what the future holds for academic networking and data transfer.

CCS CONCEPTS

• **Applied computing** → *Operations research*; • **Networks** → **Network management**; **Network monitoring**; • **Information systems** → **Data management systems**.

KEYWORDS

network performance, data transfer experience, monitoring, file transfer

ACM Reference Format:
Kathy Benninger, Chris Jordan, Michael Lambert, Lee Liming, Tabitha K. Samuel, and David Wheeler. 2022. Experiences in network and data transfer across large virtual organizations—a retrospective. In *Practice and Experience in Advanced Research Computing (PEARC '22), July 10–14, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3491418.3530763

PEARC '22, July 10–14, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9161-0/22/07.
https://doi.org/10.1145/3491418.3530763

1 INTRODUCTION

The Extreme Science and Engineering Discovery Environment (XSEDE) project was born of a need to serve the growing national community of academic researchers through a single unified interface to national academic computing resources. Following and building on the success and lessons learned through NSF's TeraGrid programs (2001–2005 and 2005–2012), the XSEDE award was funded by the NSF [17] in 2011. During the term of the first award, from 1 July 2011 to 31 December 2016, XSEDE established itself as the most advanced, powerful, and robust collection of integrated digital research resources and services in the world. XSEDE integrated and coordinated advanced digital services within the national cyberinfrastructure (CI) ecosystem to support contemporary science. This ecosystem continues to involve a highly distributed, yet integrated and coordinated, assemblage of software, supercomputers, visualization systems, storage systems, networks, portals and gateways, collections of data, instruments, and personnel with specific expertise. Due to the substantial successes of the XSEDE project, a follow-on award (often referred to as "XSEDE2" to distinguish it from the initial award) was made to continue the project in 2016 [18]. The mission of XSEDE2 was to "exist to enhance the productivity of a growing community of scholars, researchers, and engineers through access to advanced digital services that support open research by coordinating and adding value to the leading CI resources funded by the NSF and other agencies" [23]. XSEDE2 is governed by the following overarching goals:

Goals of the XSEDE2 project

- Deepen and extend use for existing and new communities through workforce development and efforts that raise awareness of the value of advanced digital services.
- Advance the Ecosystem by creating an open and evolving infrastructure and enhancing the array of technical expertise and support services.
- Sustain the Ecosystem by providing reliable, efficient and secure infrastructure, excellent CI user support services, and an innovative, effective and productive virtual organization.

XSEDE2 is organized into six different Work Breakdown Structures as shown in Figure 1 to cater to the different facets of research computing that must work together seamlessly to provide an excellent researcher experience.

The focus of this paper is a retrospective on the data transfer evolution in research computing witnessed during the lifespan of XSEDE2. Data transfer experiences within the XSEDE2 framework and its Service Providers (SPs) are facilitated by the Data Transfer Services (DTS) group which sits within the Operations area. To note, SPs are organizations with resources, funded by the NSF (or other sources), that, by formal agreement, are made part of the XSEDE community. Organizations can choose to have these resources allocated either entirely or in part through XSEDE allocation services. SPs interact with each other as well through the different facets of XSEDE, focus area specific meetings, community events, community feedback processes, and other strategic interactions.

The mission of the DTS group is to facilitate data movement and management for the community by maintaining and evolving XSEDE data services and resources. DTS plays a unique role within XSEDE in that it provides guidance on best practices within these areas even though it does not manage the data transfer endpoints. This guidance includes but is not limited to helping resources integrate into the XSEDEnet network (an L3VPN-based network, discussed in depth in section 3), providing best common practices for data transfer application deployments (both from the system administrator and researcher perspective), organizing community events such as Birds Of a Feather at conferences to discuss data and networking challenges and successes within the community, and regularly providing to stakeholders detailed information on the state of data transfer within the national academic research community.

2 BACKGROUND

In 2000, NSF began funding "terascale" systems at a number of research campuses and other sites that would "allow researchers to address problems that are too large for systems currently available." Along with awards to build and operate these unique high-performance, high-capacity computing resources, NSF issued solicitation NSF01-51 [16], Distributed Terascale Facility (DTF), with the intention of providing the large-scale funding and vision needed to build coordinated, nation-wide, distributed, terascale CI. Through NSF01-51, NSF created "an advanced, multi-site 'distributed facility' connected by ultra high-speed networking that will lead to breakthroughs and enhance the capabilities of U.S. researchers in all areas of computational, computer, and information science and engineering." This initial effort provided the infrastructure and coordination to connect four high-performance computing (HPC) sites: the National Center for Supercomputing Applications; the San Diego Supercomputer Center; Argonne National Laboratory; and the Center for Advanced Computing Research at the California Institute of Technology.

NSF funding for additional terascale resources at multiple institutions quickly led to the 2002–2004 expansion of the DTF to become the Extensible Terascale Facility (ETF), adding the Pittsburgh Supercomputing Center, Indiana and Purdue Universities, Oak Ridge National Laboratory, and the Texas Advanced Computing Center. National Center for Atmospheric Research (NCAR) resources were added in 2007. DTF and ETF programs were commonly referred to as the TeraGrid [6, 8].

As the NSF sought to further broaden the research community to include experienced and emerging groups along with non-traditional users of CI, XSEDE was announced in 2011 as the successor to TeraGrid.

2.1 Evolution of Networking and Data Transfer within XSEDE

Since the DTF era, researchers have needed to move datasets ranging in size from gigabytes to petabytes over distances of thousands of kilometers, on networks with speeds ranging from tens to hundreds of gigabits per second. Enabling this scale of data transfer, optimizing it for performance, and then making it more widely accessible to new research communities, has always been a key goal of the TeraGrid and XSEDE projects.

In the TeraGrid period, there were many notions about how researchers would need to manage their data when using national-scale systems. The community tried out many of these notions with experimental and prototypical services. We evaluated many properties of these services: individual and aggregate performance, network utilization, reliability, ease-of-use, interoperability, operational costs, etc. Notably, TeraGrid introduced Globus GridFTP servers [3] at each participating HPC site. GridFTP clients, however, were not easy to install or use, and lacked the logic necessary to maintain large-scale data transfers in the event of transient failures, common on all networks.

In 2011, XSEDE adopted Globus's software-as-a-service [4] as its primary data transfer service. The Globus web application and transfer service leverage the Globus endpoints (GridFTP servers) deployed during TeraGrid (and similar endpoints at thousands of research facilities and campuses around the world), and also incorporate numerous ease-of-use, reliability, and performance techniques to optimize and sustain transfers of all sizes in the wide area networking environment [5]. Use of the new service by XSEDE researchers began steady and rapid growth. By the time XSEDE2 began in 2016, researchers were moving large datasets to and from XSEDE's analysis systems on a daily basis. Today, the most important measures of success for data transfers in XSEDE are ease-of-use and reliability: researchers can transfer their data when necessary without a significant learning curve or technical issues.

Now, every month more than 350 individuals at over 150 institutions use XSEDE's data transfer service, Globus [9, 10], to move data from place to place. First-time users of CI account for roughly 17% of these individuals monthly, a rate that has remained constant during XSEDE2. Figure 2 shows monthly first-time, returning, and total researchers who transferred data on XSEDE2 using Globus.

Throughout 2020 and 2021, XSEDE researchers requested roughly 500 data transfers daily. The vast majority complete successfully—the most common reason for an incomplete transfer is that the researcher canceled it. Of a typical day's 500 transfers, one is likely larger than 10 TB; a half dozen are between 1 TB and 10 TB; a dozen are between 100 GB and 1 TB; and another dozen are between 50 GB and 100 GB. The remainder are less than 50 GB, and with many less

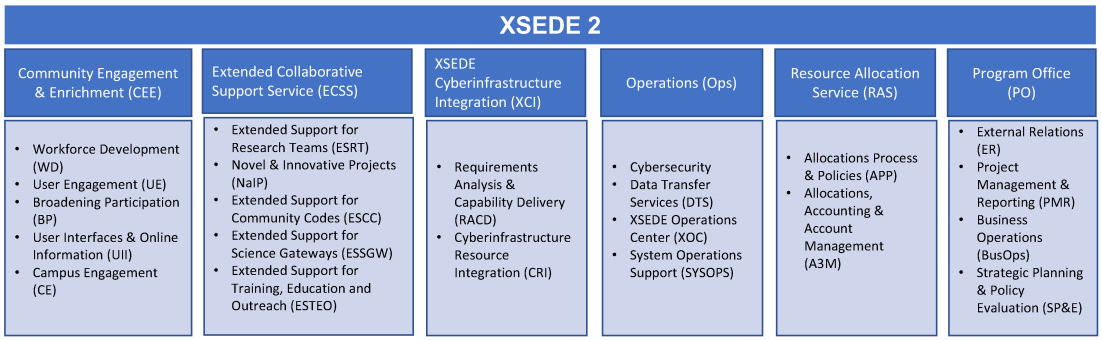

Figure 1: XSEDE2 Work Breakdown Structure

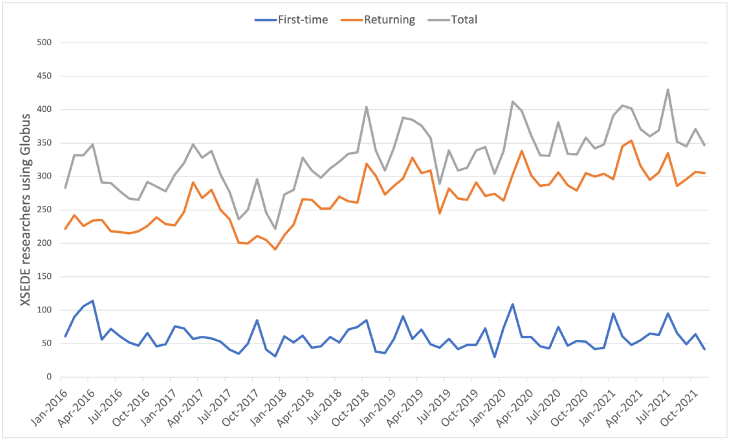

Figure 2: Monthly number of researchers transferring data on XSEDE2 with Globus

than 1 GB. Figure 3 shows the annual number of XSEDE2 data transfers of various sizes, from 100 GB to more than 10 TB. In each range, there has been constant and growing use throughout XSEDE2. We also see even larger transfers with increasing frequency, such as for large computations [15] or when a storage provider or research facility is decommissioned [1].

XSEDE researchers transfer data between hundreds of research storage systems. Data transfers are typically: (a) between a university or lab system and an XSEDE HPC resource (in either direction), or (b) between a pair of HPC resources. Transfers larger than 1 TB are almost exclusively in the latter category. While many U.S. campuses have research IT personnel who maintain systems for use with XSEDE, XSEDE also serves researchers at smaller universities and colleges that do not provide such staffing resources. Thus XSEDE's data transfer service has to be easy for researchers to use across a variety of systems: Windows, Macintosh, and Linux computers; Windows and Linux cloud systems.

3 NETWORKING EXPERIENCES

While TeraGrid and XSEDE have evolved and undergone significant growth over the past two decades, meeting the fundamental need for high-performance networking and data transfer has remained a priority. The high-performance network infrastructure funded by the NSF was a key component in making this nation-wide distributed CI resource possible. Throughout the project, network engineers at the participating sites have collaborated to design the infrastructure that would support the research community's needs. Part of the challenge in developing TeraGrid, one of the first dedicated academic R&E networks, was the innovative design decisions that needed to be made. As the underlying technologies have changed, XSEDE network engineers still seek optimal solutions to many of the same types of issues.

3.1 Wide area R&E networking

Throughout the progression of projects from DTF to XSEDE2, a number of providers and technologies have been used for the primary data transport paths between end sites[1]. To a large extent, this is a reflection of the evolution of research and education networks across the United States. Since the transition from TeraGrid to XSEDE, we have referred to this interconnection network, somewhat informally, as XSEDEnet.

During the DTF era into the ETF era, sites were connected with dedicated 10 Gb/s circuits leased from Qwest[2]. These circuits were provisioned as a mixture of OC-192 SONET, 10 GbE WAN PHY and 10 GbE LAN PHY. Late in the TeraGrid project, extending into the early years of XSEDE, connectivity was migrated from Qwest to National LambdaRail (NLR), still at 10 Gb/s. SONET connections

[1]It should be noted that other connectivity between sites, for example over research and education networks, is assumed, if not mandated.
[2]Individual links were 10 Gb/s, but backbone connections and some connections to sites were aggregated. This allowed bandwidths greater than 10 Gb/s, but limited any single-stream data transfer to 10 Gb/s.

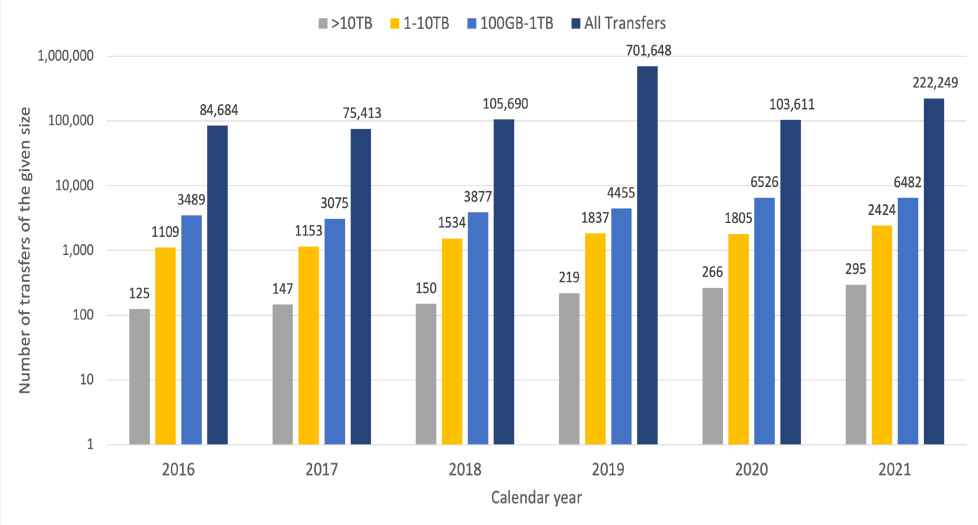

Figure 3: Large data transfers during XSEDE2. Note log scale in y-axis.

were phased out to reflect the Ethernet-based layer-two backbone (FrameNet) of NLR. At the outset of DTF, 10 Gb/s was at the leading edge of bandwidth to academic sites; throughout TeraGrid and by the full transition to NLR in 2011, 10 Gb/s connectivity was becoming accessible and increasingly used for university R&E connections. These enhanced bandwidths enabled data transfers at much greater rates than would have been possible via normal campus connectivity.

Following the transition of NLR control from the academic community to a private individual [7], what we now call XSEDEnet migrated to Internet2 for transport. Initial 10 Gb/s connections have since all been upgraded to at least 100 Gb/s, and XSEDEnet bandwidth is now a component of each site's Internet2 connection rather than being provisioned as a separate circuit.

3.2 Architecture and networking services

XSEDEnet has undergone several changes in architecture since its inception with DTF. Through the TeraGrid era, it was purely a routed layer-three network. Sites connected through NSF-subsidized backhaul circuits to routers in Chicago and Los Angeles and peered with those routers by using BGP, with IPv6 routing from July 2003 forward. With the migration to Internet2 transport, XSEDEnet replaced dedicated circuits with a layer-two virtual network. Initially, this was provisioned using Internet2's Open Exchange Software Suite (OESS) tool [2] as a multi-point network with BGP route servers for peering. Limitations of OpenFlow 1.0 added operational complexity because it was necessary to enter next-hop MAC addresses manually. Following the retirement of OpenFlow from the Internet2 backbone, XSEDEnet was reprovisioned as a VLAN mesh with pairwise connections (and BGP peering sessions) between all sites, still using OESS for management.

As the number of XSEDE SPs grew, DTS engineers and CI engineers at the new sites realized that the n^2 scaling issue inherent in expanding a point-to-point VLAN mesh network design would quickly make the architecture unwieldy and unsustainable. The group thus asked Internet2 management about the possibility of

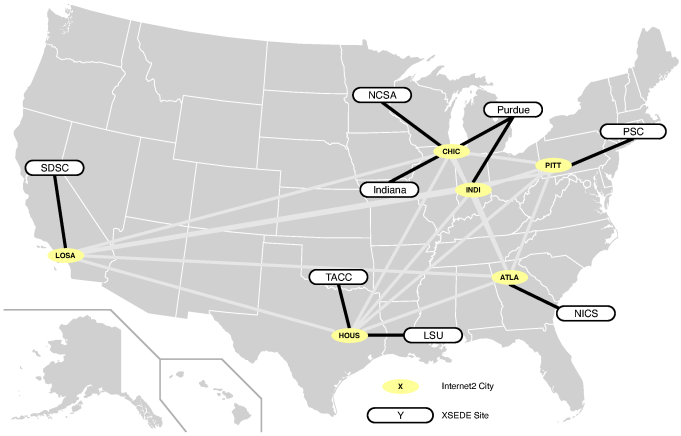

Figure 4: Schematic of XSEDEnet as of February, 2022

replacing the VLAN mesh with a virtual private network (VPN) overlay, either at layer-two (probably using BGP peering between sites and route servers) or at layer-three (BGP peering between sites and Internet2 routers). Internet2 was amenable to an L3VPN. Migration to this VPN led to the current XSEDEnet architecture, shown in Figure 4. A site wishing to participate in the network provisions a VLAN to the nearest Internet2 router, and the Internet2 NOC assigns IP addresses for the VLAN and coordinates the turn-up of peering between the site's router and the XSEDE virtual-routing instance on the Internet2 router.

Looking beyond the state of the practice for high-performance networking, TeraGrid and XSEDE served as test platforms for two different experiments in programmable network bandwidth provisioning. The goal of each experimental infrastructure was to optimize the use of limited bandwidth resources by dynamically allocating bandwidth as needed. The first service was Sherpa [22], offered in 2009 by NLR. The second was an NSF CC∗ funded project undertaken by XSEDE engineers in 2014–2016, the OpenFlow-based

DANCES [14]. Both projects successfully demonstrated their intended capabilities but were ultimately not widely adopted due to implementation complexity (both) or superseded by the order-of-magnitude increase in wide area connectivity to 100 Gb/s (DANCES).

The only network service supported on XSEDEnet at this time is layer-3 connectivity to the other participating sites; it is up to sites to work together to provision layer-2 connectivity using OESS. The assumption is that a site will make its XSEDE-allocated resources (and typically its Science DMZ) available over XSEDEnet.

3.3 Network monitoring and performance measurement

As affordable R&E wide area connection speeds have evolved toward 10 Gb/s, 100 Gb/s, and even 400 Gb/s, end sites no longer need to install and maintain dedicated, special purpose circuits for connecting their high-performance CI resources. Monitoring and maintaining connectivity has become part of standard production network support for XSEDE sites. XSEDE relies on its backbone provider organizations, Internet2 and GlobalNOC, for ongoing routine support and quarterly reports to provide insight into network operations, troubleshooting, and performance trends.

End-to-end performance measurement and improvement have been a priority since the beginning of XSEDE. Network performance measurement on TeraGrid was done with 1 GbE-connected "netmons", custom-built by TeraGrid engineers to run basic network tests for bandwidth, loss, and latency. By the time of XSEDE, the perfSONAR[13] network measurement platform was rapidly being adopted within the R&E networking community. In 2012, DTS deployed 10 Gb/s-connected perfSONARs at the eight Level-1 SPs for network measurement and testing among these core XSEDE sites. Since that time, perfSONAR has become the standard tool for network performance testing in the R&E community to the degree that NSF CI solicitations frequently state that "Proposals are expected to describe an approach to end-to-end network performance measurement based on the perfSONAR framework with associated tool installation and use; proposals may describe an alternative approach to perfSONAR with sufficient justification." [19] In 2019, DTS deployed four 100 Gb/s-connected perfSONARs. Testing with 100 Gb/s has had limited usefulness, with meaningful testing requiring bandwidth/buffer control and tweaking.

perfSONAR provides several valuable testing capabilities in both a periodic, automated scheduling mode and a manual, on-demand mode. These tests include throughput, loss, latency, and traceroute. Another important capability of perfSONAR is the ability to initiate tests between sites without requiring a login at the remote site. perfSONAR test results can be displayed graphically to give a ready view into current conditions and past trends.

3.4 Takeaways

The network evolution and operation of TeraGrid and XSEDE have provided several valuable insights and takeaways. Foremost in operational experience is that end-to-end high-performance network infrastructure is still subject to the negative impacts of common "last-mile" problems within the end site's network. These problems are most commonly due to inadequate packet buffering within the CI or applications, broken path MTU discovery, or incorrect routing rules or policies. Fortunately, the latter two issues can be identified through use of commonly available network diagnostic tools such as ping, traceroute, and tracepath and then corrected with a software configuration change. Unfortunately, buffer constraints are more difficult to diagnose, requiring familiarity with data transfer applications and their settings, TCP performance optimization, router/switch queuing, and the underlying buffer hardware design of the routers/switches. Some of these issues can be resolved by software configuration, but an inappropriate router/switch choice can only be remedied by hardware replacement. To disseminate this information to the community, DTS has offered one-on-one site consulting and has presented BoFs at PEARC.

In the architecture of TeraGrid, the multiple aggregated links demonstrated the debugging complexity and inconsistent throughput that can result from the failure of a link in an aggregated link set. Aggregated links were also a source of confusion in network throughput expectations as researchers saw a 40 Gb/s connection comprising 4×10 Gb/s links but their individual data transfers never achieved more than 10 Gb/s.

From experience on both TeraGrid and XSEDE, overprovisioning of network bandwidth offers a significant advantage in operational simplicity over mechanisms for dynamic bandwidth allocation and management.

4 DATA TRANSFER

Over the 20+ years since the beginning of TeraGrid, some data transfer technologies and goals have been remarkably consistent, while the overall character of the data transfer infrastructure, and the size and nature of the community it serves, have changed significantly. We discuss here some of the technologies used to support data transfer over the high-performance TeraGrid and XSEDE networks, the mechanisms used to measure and monitor those transfers, and the definitions used to determine whether this infrastructure was succeeding at serving the targeted communities.

4.1 Data Transfer Technologies

The "GridFTP" client and server, and later the Globus management interface to these servers, have been components of the XSEDE architecture since the early TeraGrid era, with little change in the core server technology and data transfer protocols. Much of the change in data transfer administration over this period has been in configuration changes, practices around service and usage monitoring, and the relative importance of various mechanisms for researchers to access the GridFTP servers deployed at various TeraGrid and XSEDE sites.

The end-to-end performance achieved over DTF, TeraGrid, and XSEDEnet over time has benefited from sustained technological innovations in several areas. The wide area networks got faster. Equally important for the "last mile" was the adoption of ScienceDMZ concepts, the deployment of carefully configured Data Transfer Nodes (DTNs) [12], and the adoption of the cloud-hosted Globus service and associated design patterns [11] to configure and drive transfers among pairs of such endpoints.

Wide-area file systems were experimented with substantially in TeraGrid, and were reasonably successful in terms of basic functionality, stability, and performance given the right workloads. However,

challenges around the funding model and appropriate allocations procedures ultimately made these technologies too difficult to maintain by the time of XSEDE. This is an example of a mismatch of CI implementation and CI practice in the target community, as individual projects with appropriate needs were able to make successful use of these deployments in demonstration cases, but we were not successful at making these technologies a long-term production component.

There have also been a number of tools developed and/or supported within the project over the years to assist researchers in managing their access to data transfer and other software tools, and to assist the project in monitoring and deploying these tools. At the time of writing, however, and for most of the lifetime of XSEDE, most data transfer activity within the project, in raw byte count and number of files transferred, is performed via Globus GridFTP servers and managed through the Globus interface. All XSEDE resources also support the use of SSH/SCP for data transfer, with High-Performance Networking extensions[21], and this facility is routinely used for transfer of small data sets, but insofar as XSEDE is focused on enabling large-scale data transfer, and optimal use of high-performance networks, activity is currently focused around the Globus/GridFTP toolset.

4.2 Goals of Data Transfer

There are two key dimensions around which data transfer within the XSEDE context, and the efforts within the project to measure and improve transfer activities, have been organized. The first, and most important during the early iterations of the project, is raw performance and volume of data transferred. Particularly during the early TeraGrid era, when national-scale dedicated network links were a unique and valuable aspect of the overall project, having data transfer tools that could effectively take advantage of these capabilities was a major goal, and everything from software tool selection to configuration and documentation was centered around these performance aspects.

These changes also track with the changing character of the project over time, from the TeraGrid era when the project encompassed four to five large sites with high performance machines, to more recent years when XSEDE includes around 30 SPs at any given time. As XSEDE has become more diverse along multiple dimensions, the priority of enabling effective use by a broader CI community has risen relative to the more technically focused, performance-centric early efforts. These efforts have primarily focused on documentation, consultation on operating similar software stacks and otherwise promoting interoperability with the data transfer infrastructure of XSEDE.

4.3 Measuring Performance

In the early iterations of the project, while low-level network monitoring was supported, little direct information specific to data transfer activity was available. While broad conclusions could be reached about overall network utilization for data transfers and the ability of data transfer utilities to use all available bandwidth, more specific information about data transfer characteristics (e.g., file sizes and counts) and the performance achieved at specific endpoints was lacking.

Early in the course of the XSEDE program, an effort was undertaken to develop tools to store and parse GridFTP logs, store the resulting output within a database, and query the database to generate high-level statistics regarding, for example, the number of files transferred to and from a selected resource within a given date range. This transfer logging also provided sufficient detail to generate some basic performance data and added a significant new technical capability to the XSEDE toolset.

However, the distributed nature of XSEDE and the differing goals and policies of member institutions led to significant challenges in gathering the required log data. Logs needed to be gathered regularly for each resource, but some sites determined that the amount of information included in the logs was more than they wished to share, necessitating in some cases additional scrubbing of log data, and in other cases a total lack of data for certain resources. Additionally, as more resources were added, the number of records became large, as XSEDE endpoints collectively participate in tens of millions of individual file transfer operations each year. These and other concerns limited the utility of the overall logging system, which remained at best a tool to provide high-level information regarding transfer volume, overall performance, and other aggregate properties.

The capability most used in recent years to track information on data transfers is the Globus centralized statistics, which take advantage of the fact that the Globus web interface for controlling data tracking has increased in popularity over the years. As a centralized resource, Globus can record information on each transfer initiated and controlled through the Globus web interfaces. While this does potentially miss those transfers which are initiated through mechanisms outside the Globus web interfaces, at the present time these constitute a small minority of the total volume of data transfers, as the web interface has become the primary recommended tool for managing data transfers within XSEDE. This tool is useful for tracking the extent to which data transfer activity is growing or shrinking, identifying large-scale changes in the nature of transfer activity, and assessing the relative importance of data moving between resources within XSEDE vs moving between institutional resources and XSEDE resources.

Another monitoring resource that was useful in understanding the practical limitations of pairwise data transfers between XSEDE GridFTP endpoints was the "Speedpage" developed at the Pittsburgh Supercomputing Center [20]. This application twice daily performed a set sequence of pairwise transfers between such endpoints, using both memory-to-memory and disk-to-disk modes, and displayed results in a public web page, providing system administrators and researchers with a view of the functional status, and potential performance, of any given endpoint pair. Replacing the Speedpage would be of benefit to the community.

4.4 Defining Success

A recurring topic of discussion within XSEDE over the years has been the question of how to measure "success" for Data Transfer. In early years, success was defined simply in terms of enabling a certain level of performance, fully utilizing the network and storage capabilities at either end of given resource pairs. Over time, there was need for specific measurements of what optimal performance

meant, and to ensure that researchers regularly achieved theoretical peak performance levels that systems were capable of achieving. The numbers of data transfer practitioners, and the numbers of transfers performed, emerged as additional metrics for success.

For example, in the early TeraGrid years, significant effort was put into assessing the optimal sets of transfer parameters for GridFTP transfers, and into ensuring that network stacks were optimized for transfer of large data sets over high-bandwidth, high-latency links. Definitions of successful data transfer operations were consequently oriented around achieving certain levels of both data transfer performance and network utilization. As the file transfer mix became more diverse and the number of CI users increased, it became more difficult to make assumptions about the conditions of file transfer and the knowledge of these CI users.

More recently, as the infrastructure has become a robust production operation with a stable software base, and network capacity has become more plentiful, definitions of success have become oriented towards new and total researcher counts and overall transfer volumes. These new metrics reflect the increased importance of ensuring that a broad research community makes effective use of stable infrastructure, rather than just optimizing the network and data transfer components for peak performance. It also reflects the transition from a new, more experimental architecture in the early TeraGrid years to the long-term production infrastructure that is being operated today.

5 TAKEAWAYS

The national research community and consequentially the XSEDE user base have grown exponentially over the past decade. More than 50 centers and resources have joined the XSEDE consortium since the beginning of XSEDE2, and the increase in the data volumes transferred over this time frame strongly reflects this growth. In the following we highlight some takeaways relating to data movement and networking that the group has gathered from its years of operations.

5.1 Implementation and practice change with a changing landscape

The initial vision of the TeraGrid project was to expand access to NSF's expensive, high-performance computing resources to a much larger community of researchers than could be effectively supported by the individual sites. In NSF's words, "The goal of this solicitation is to achieve the most computational infrastructure for the broadest scientific and engineering community within the funds available" [16]. The TeraGrid project brought together a small number of principal sites that helped facilitate academic research for the nation. Today, XSEDE2 has over 40 partner institutions. This growth has been made possible by the work of multi-site collaborative teams who have selected, and in some cases created, the best practices, tools, and applications required to sustain a stable, leading-edge, computational environment for a broad range of scientists and engineers.

Increasing numbers of Service Providers have required enhanced coordination between sites to enable seamless user access for data movement across all facilities. The issues that need to be addressed for effective data transfer include several fundamental infrastructure and services such as: network connectivity; data transfer compatibility, implying interoperable data transfer applications and endpoints; and shared user identity and access management across sites. Although the capabilities of the CI and its services have advanced significantly over time, the fundamental needs have remained. In TeraGrid and XSEDE, Networking and Data were separate teams. With the start of XSEDE2 in 2016, the Networking and Data groups were merged to form the Data Transfer Services (DTS) team. Because of the strong interdependence of network performance and efficient data movement, combining these groups has enabled direct collaboration between these two groups and enhanced XSEDE2's ability to effectively respond to user needs.

5.2 Need an increased focus on communication, engagement and documentation

With high throughput and a high volume of data transfers becoming commonplace in academic institutions across the nation, the need for communication, engagement, and documentation of data transfer practices is greater than ever. Sharing of best practices, how-tos, and challenges in setting up and maintaining highly reliable networks and data transfer endpoints will be highly beneficial to emerging communities of research. Stronger communication channels between universities can also help in securing CI. As researchers use resources spread across the nation, and are not restricted to their own institutional resources, sharing of user experiences and obstacles to delivering science will help reduce costs that would be incurred if each institution had to navigate these obstacles alone.

5.3 Standardization of tools across campuses

Unlike most other system administrator tools and services, where there are a plethora of choices available, there has been a standardization around Globus as the service of choice for providing highly reliable, high throughput data transfers across the nation's campuses. OS-native options such as scp, sftp (optimally, both of those tools with HPN-SSH modifications), and rsync are still used, but their poor performance discourages broad adoption.

5.4 Challenges remain in extending support to the community

The DTS group added a focus survey section to the annual XSEDE user surveys in 2019. In 2020, it undertook an effort to advertise its availability to help individual researchers solve their data transfer challenges, and also to get a better understanding of the current landscape of data transfer performances from an end researcher point of view. While the response rate to the survey was satisfactory, and several users indicated they would like to receive DTS support in solving their data transfer issues, the engagement after this survey was disappointing. There was minimal response from survey respondents when the group tried to reach out to them to investigate their data transfer issues. This could be attributed to either the researcher having moved on to other research projects or resources that negated the need for extended support, or that the researcher had alleviated the issue through support obtained from their local institution.

5.5 Setting realistic user expectations for data transfer performance is important

Often researchers are curious as to why their data transfers performances are not as fast as their campus networks would allow them to be. It is important for system administrators and other campus research facilitators to educate users on the actuality that any given data transfer can only perform as well as the slowest component involved in the transfer. It may be a single slow network link, the disk drive on a researcher's laptop, or an overloaded file system on a supercomputer. Regardless, the overall transfer can perform no faster than the slowest component in the chain. It is vital for campus research facilitators to set realistic expectations for data transfer performance and, more importantly, help researchers investigate and resolve any intermediate points that might impede data transfers.

5.6 Wider Implications

Virtual CI organizations must establish a solid, shared understanding of the goals of both their project and of the technical CI, and how these components serve the goals of the project and the target communities. The organization must remain adaptable to these goals, and to the changing technical and infrastructural landscape, over long periods of time, without changing course constantly and confusing one's user community. Over time, the community will develop expectations of the CI, and consistently meeting those expectations will become a central aspect of what the organization does. This, in and of itself, should be viewed as success—any infrastructure component doing its job should become all but invisible until it goes away.

6 CONCLUSION

XSEDE began as a collaborative project of multiple institutions and has expanded to impact hundreds of institutions over more than ten years of operation. Over this time, the XSEDE community has grown to support many thousands of users of CI at all career stages. Key to this impact is the tremendous, successful support from the Data Transfer Services team. By providing well monitored and maintained data transport services, including the various software and hardware components that support effective end-to-end data movement, the DTS group has enabled ever increasing scientific discovery. This effort focused on gathering researchers' needs and on the long-term goal of improving the overall researcher experience with data transfers. This was part of a strategy for collecting best practices and advancing understanding throughout the community of how best to design and implement storage and network systems for end-to-end data movement and sharing these practices to the XSEDE community and beyond.

ACKNOWLEDGMENTS

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), supported by National Science Foundation grant number ACI-1548562. We also thank the XSEDE senior management team, current and former members of the XSEDE Data Transfer Services, Internet2, and Globus teams for their support and contributions.

REFERENCES

[1] 2021. https://news.utexas.edu/2021/05/10/priceless-astronomy-data-saved-after-collapse-of-arecibo-telescope/.
[2] Internet 2. 2016. OESS. https://internet2.edu/network/services-for-network-providers/oess-open-exchange-software-suite
[3] William Allcock et al. 2005. The Globus striped GridFTP framework and server. In *ACM/IEEE Conference on Supercomputing (SC '05)*. 54.
[4] Bryce Allen et al. 2012. Software as a service for data scientists. *Commun. ACM* 55, 2 (2012), 81–88.
[5] Rachana Ananthakrishnan et al. 2015. Globus platform-as-a-service for collaborative science applications. *Concurrency and Computation: Practice and Experience* 27, 2 (2015), 290–305.
[6] Peter H Beckman. 2005. Building the TeraGrid. *Philosophical Transactions of the Royal Society A* 363, 1833 (2005), 1715–1728.
[7] businesswire.com. 2011. https://www.businesswire.com/news/home/20111116005513/en/National-LambdaRail-Announces-Expanded-Partnership-with-Cisco-Systems
[8] Charlie Catlett et al. 2008. TeraGrid: Analysis of organization, system architecture, and middleware enabling new types of applications. In *High Performance Computing and Grids in Action*. IOS press.
[9] Kyle Chard et al. 2014. Efficient and secure transfer, synchronization, and sharing of big data. *IEEE Cloud Computing* 1, 3 (2014), 46–55.
[10] Kyle Chard et al. 2016. Globus: Recent enhancements and future plans. In *XSEDE16*. Article 27, 8 pages.
[11] Kyle Chard et al. 2018. The Modern Research Data Portal: A design pattern for networked, data-intensive science. *PeerJ Comput Sci* 4 (2018), e144.
[12] Eli Dart et al. 2014. The science DMZ: A network design pattern for data-intensive science. *Scientific Programming* 22, 2 (2014), 173–185.
[13] Andreas Hanemann et al. 2005. PerfSONAR: A service oriented architecture for multi-domain network monitoring. In *3rd International Conference on Service-Oriented Computing*. 241–254.
[14] Victor Hazlewood et al. 2016. Developing applications with networking capabilities via end-to-end SDN (DANCES). In *XSEDE16*. 1–7.
[15] Rajkumar Kettimuthu et al. 2018. Transferring a petabyte in a day. *Future Generation Computer Systems* 88 (2018), 191–198.
[16] NSF. 2001. Distributed Terascale Facility (DTF). https://www.nsf.gov/pubs/2001/nsf0151/nsf0151.htm. Accessed: 2022-02-15.
[17] NSF. 2011. https://nsf.gov/awardsearch/showAward?AWD_ID=1053575
[18] NSF. 2016. https://www.nsf.gov/awardsearch/showAward?AWD_ID=1548562
[19] NSF. 2021. https://www.nsf.gov/pubs/2021/nsf21528/nsf21528.htm. Accessed: 2022-02-15.
[20] PSC. 2020. PSC Speedpage. https://confluence.xsede.org/download/attachments/2491275/Speedpage_Sept2016.pdf. Accessed: 2022-02-15.
[21] Chris Rapier et al. 2021. HPN-SSH. https://www.psc.edu/hpn-ssh-home/. Accessed: 2022-05-25.
[22] Global research network operating center. 2020. NLR dynamic VLAN services & the Sherpa provisioning tool. https://www.nitrd.gov/subcommittee/lsn/jet/material/DVS_Presentation_for_JET_031709.pdf. Accessed: 2022-02-15.
[23] XSEDE. 2016. XSEDE Governance. https://www.xsede.org/about/governance. Accessed: 2022-02-15.