Encyclopedia of Information Science and Technology, Fourth Edition

Mehdi Khosrow-Pour
Information Resources Management Association, USA

IGI Global
DISSEMINATOR OF KNOWLEDGE

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Khosrow-Pour, Mehdi, 1951- editor.
Title: Encyclopedia of information science and technology / Mehdi Khosrow-Pour, editor.
Description: Fourth edition. | Hershey, PA : Information Science Reference, [2018] | Includes bibliographical references and index.
Identifiers: LCCN 2017000834| ISBN 9781522522553 (set : hardcover) | ISBN 9781522522560 (ebook)
Subjects: LCSH: Information science--Encyclopedias. | Information technology--Encyclopedias.
Classification: LCC Z1006 .E566 2018 | DDC 020.3--dc23 LC record available at https://lccn.loc.gov/2017000834

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Cyberinfrastructure, Cloud Computing, Science Gateways, Visualization, and Cyberinfrastructure Ease of Use

C

Craig A. Stewart
Indiana University, USA

Richard Knepper
Indiana University, USA

Matthew R. Link
Indiana University, USA

Marlon Pierce
Indiana University, USA

Eric Wernert
Indiana University, USA

Nancy Wilkins-Diehr
San Diego Supercomputer Center, USA

INTRODUCTION

Computers accelerate our ability to achieve scientific breakthroughs. As technology evolves and new research needs come to light, the role for cyberinfrastructure as "knowledge" infrastructure continues to expand. In essence, cyberinfrastructure can be thought of as the integration of supercomputers, data resources, visualization, and people that extends the impact and utility of information technology. This article defines and discusses cyberinfrastructure and the related topics of science gateways and campus bridging, identifies future challenges in cyberinfrastructure, and discusses challenges and opportunities related to the evolution of cyberinfrastructure and cloud computing.

BACKGROUND

Today's US national cyberinfrastructure ecosystem grew out from the National Science Foundation-funded supercomputer centers program of the 1980s (National Science Foundation, 2006). Four centers provided supercomputers and support for their use by the US research community. Researchers generally accessed one supercomputer at a time, sometimes logging into a front-end interface. At this time, the focus of the research computing community was centered on supercomputers – traditionally defined as computers that are among the fastest in existence. Over time there have been several different supercomputer architectures, but the key points were that supercomputers were monolithic systems that were among the fastest in the world. At present we can think of supercomputers as being a subset of the more general term high performance computer (HPC) – where HPC means that many computer processors work together, in concert, to solve large computational

DOI: 10.4018/978-1-5225-2255-3.ch092

challenges and where the computer processors communicate via very fast, networks internal to the HPC system. HPC focuses on computing problems where a high degree of communication is needed among the processors working together on a particular problem. HPC is a more general term than supercomputers because there are many HPC systems that are modest in total processing capacity relative to the fastest supercomputers in the world (cf. Top500.Org, 2016).

In the early days of supercomputing, using multiple supercomputers in concert was not possible. In the late 1980s, the National Research and Education Network initiative created several testbeds for distributed computing, including the CASA testbed which linked geographically distributed supercomputers to solve large-scale scientific challenges (U.S. Congress Office of Technology Assessment, 1993). A turning point in distributed high performance computing was the I-WAY project – a short-term demonstration of innovative science enabled by linking multiple supercomputers with high performance networks (Korab & Brown, 1995). It demonstrated the possibilities to advance science and engineering by linking supercomputers using high-speed networks.

In the late 1990s, the NASA Information Power Grid provided a production grid of multiple supercomputers connected by a high-speed network (Johnston, Gannon, & Nitzberg, 1999). Around this time began also the concept of high throughput computing (HTC) with a software system called Condor (Litzkow, Livny, & Mutka, 1988). HTC takes the approach of breaking a problem up into small pieces of work and distributing them to multiple CPUs over network connections that may be relatively slow. HTC best suits problems where relatively little communication is needed among the processors working together on a particular problem or simulation. Because HTC applications can operate relatively efficiently on processors with little communication among the processors, HTC applications have always fit naturally into a distributed computing environment (Thain, Tannenbaum, & Livny, 2005). Today, a popular framework for distributed storage and processing of large data sets is Apache Hadoop (The Apache Software Foundation, 2006).

Over time, distributed computing evolved into 'grids,' with grids emerging as a commonly used term in the late 1990s. Typically, computational grids are the hardware and software infrastructure which provides access to the computational capabilities (Foster & Kesselman, 1998, 2004). *Middleware* is a key software component of cyberinfrastructure, enabling the disparate components of cyberinfrastructure to work together. In effect, middleware manages complex interactions between resources which allows for the development of new networked applications (National Science Foundation, 2004). Around the turn of the century, the US government funded two major grid projects – TeraGrid and the Open Science Grid. In 2001, the NSF funded an experimental computational, storage, and visualization resource called TeraGrid, which developed grid capabilities for supercomputer centers (National Science Foundation, 2006). The Open Science Grid (OSG) (Livny et al., 2006; Open Science Grid, 2015), first funded with that name in 2006, grew out of three projects that developed HTC grids for the purpose of analyzing data from physics experiments (Avery, 2007).

Tying geographically distributed computing systems together into grids to create a whole greater than the sum of its parts was widespread around the turn of the century. However, the term grid computing was becoming laden with sometimes competing definitions. In addition to computing and data grids, other terms such as collaboration, semantic, and peer-to-peer grids emerged, distinguished by the characteristics of the protocols and interactions between components (Fox, 2006). The potential for confusion and competing definitions of different types of grids led Dr. Ruzena Bajcsy, then NSF assistant director of the Computer and Information Science and Engineering Directorate, to use the term cyberinfrastructure when charging a new advisory group to offer advice to

the NSF – the "Blue Ribbon Advisory Panel on Cyberinfrastructure." The term cyberinfrastructure had been used before, in a different sense, by Richard Clarke, then US National Coordinator for Security, Infrastructure Protection, and Counterterrorism (Clarke & Hunker, 1998). Bajcsy stated that she used the term cyberinfrastructure because she wished to "create a program … that would involve the broader computer science/information technology community" (Bajcsy, 2013). The committee report goes on to state, "the newer term cyberinfrastructure refers to infrastructure based upon distributed computer, information and communication technology. If infrastructure is required for an industrial economy, then we could say that cyberinfrastructure is required for a knowledge economy" (Atkins et al., 2003).

Bajcsy's successor at the NSF, Dr. Peter Freemen, stated that this report "led to the creation of a term for infrastructure that attempts to capture the integration of computing, communications, and information for the support of other activities (especially scientific in the case of NSF)" (Freeman, 2013). In 2007, Freemen wrote "cyberinfrastructure can have many definitions and, to some extent, the definition is in the eye of the beholder" (Freeman, 2007). To make it clearer for scientists outside of science and physics, Indiana University developed a definition identifying components and function:

Cyberinfrastructure consists of computing systems, data storage systems, advanced instruments and data repositories, visualization environments, and people, all linked together by software and high performance networks to improve research productivity and enable breakthroughs not otherwise possible. (Stewart et al., 2010)

The EDUCAUSE Campus Cyberinfrastructure Working Group and the Coalition for Academic Scientific Computation developed a definition which includes teaching and learning:

Cyberinfrastructure consists of computational systems, data and information management, advanced instruments, visualization environments, and people, all linked together by software and advanced networks to improve scholarly productivity and enable knowledge breakthroughs and discoveries not otherwise possible. (Dreher et al., 2009)

The characteristics distinguishing cyberinfrastructure from other IT terms and concepts is the inclusion of resources like instruments and sensor networks as well as people and a focus on knowledge breakthroughs. Cyberinfrastructure may be distinguished in particular from the more European term eScience on the basis of the explicit role of people in cyberinfrastructure. eScience is defined as "the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualization back to the individual user scientists" (National e-Science Centre, 2010).

CYBERINFRASTRUCTURE TODAY

The broad use of cyberinfrastructure in science and engineering envisaged by Bajcsy is in ample evidence today. That cyberinfrastructure enables breakthroughs not otherwise possible is demonstrated by two Nobel prizes for work made possible by major cyberinfrastructure resources – the Open Science Grid and XSEDE.

The Open Science Grid is an international HTC resource. Many different organizations own the computers participating in the grid (Open Science Grid, 2015). OSG's people part of cyberinfrastructure is organized through dozens of Virtual Organizations (VOs) that use the computational resources of the OSG, each supporting its own uses and users. Analysis of data from the Large

Hadron Collider (LHC) is the paradigmatic use case for HTC. LHC data can be broken down into large numbers of small data sets, each of which may be analyzed in isolation. The 2013 Nobel Prize for Physics was awarded to François Englert and Peter Higgs for the theoretical discovery of the particle now known as the Higgs Boson. The existence of the Higgs Boson was verified in experiments at the LHC, with the data analyses made possible by the OSG.

The largest HPC-oriented cyberinfrastructure in the use is the eXtreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014). XSEDE is a single, virtual system which is comprised of a collection of integrated and highly-advanced digital resources and constitutes the largest HPC resource funded by the US government (Towns et al., 2014; XSEDE, 2013, 2015). The 2013 Nobel Prize in Chemistry was awarded to Martin Karplus, Michael Levitt and Arieh Warshel, for the development of multiscale computer models of complex chemical systems. Karplus used resources of the TeraGrid, the predecessor of XSEDE, and Warshel uses resources of XSEDE (XSEDE, 2015).

Cyberinfrastructure may support a particular research domain or application. Cyberinfrastructure has also been widely adopted in the private sector, particularly in advanced engineering, medicine and pharmaceuticals, mining and oil exploration, finance, and manufacturing (Tabor Griffin Communications, 1998).

Cyberinfrastructure systems need not be massive to be important. A Specialized cyberinfrastructure supports NASA's *Operation IceBridge* in measuring polar ice sheets in Greenland and Antarctica. *Operation IceBridge* uses sophisticated synthetic aperture radar (SAR) systems to study polar ice and map the bedrock base in Greenland and Antarctica (Hayden, Fox, & Gogineni, 2007; Knepper, Link, & Standish, 2015). One of the characteristics of SAR is that one doesn't get an image out of SAR systems directly; a great deal of computation is required to generate an image. In-plane computation and data storage provide real-time analysis of multiple radar data sources (Figure 2). This cyberinfrastructure is highly specialized to deal with the rigors of fieldwork in Antarctica. The cyberinfrastructure designed to support *Operation IceBridge* enables real-time

Figure 1. The open science grid

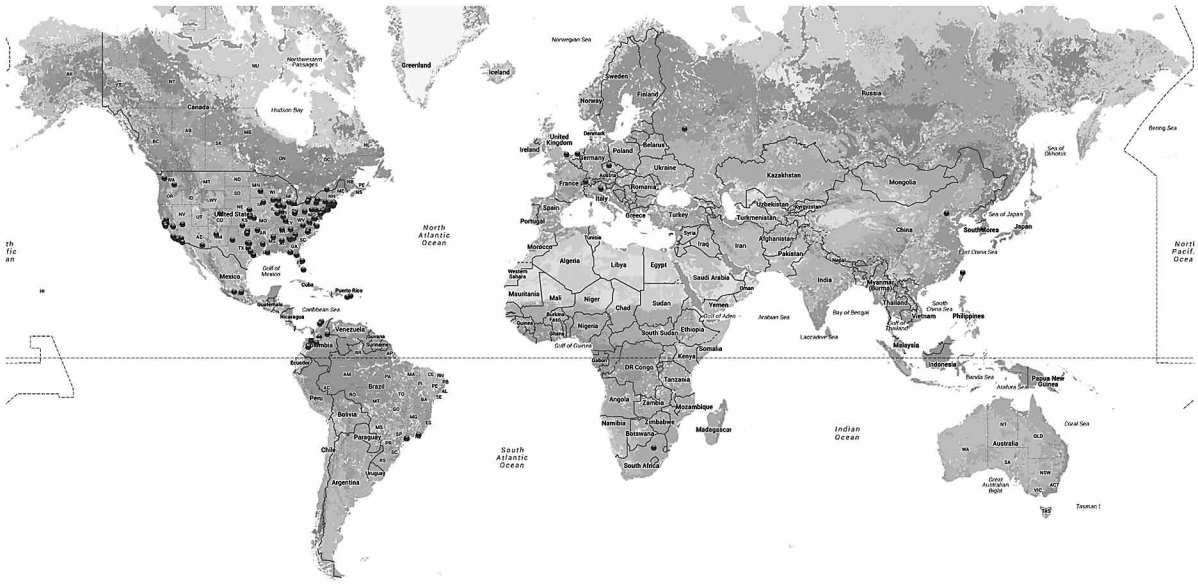
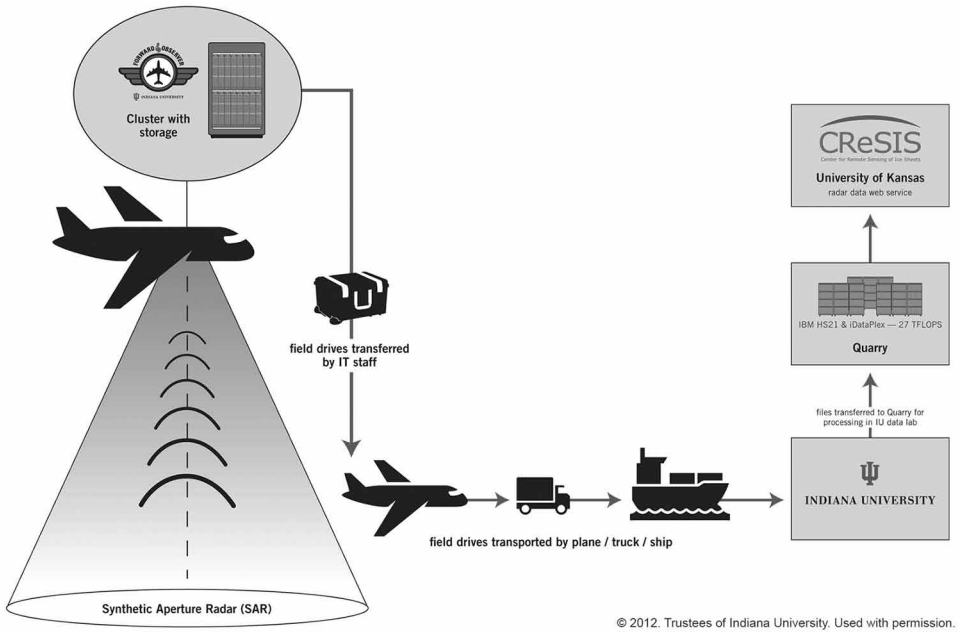


Figure 2. NASA operation icebridge field radar data processing cyberinfrastructure
Source: Knepper, Standish, & Link 2015



reactions to data as it is being collected in an airplane over the Antarctic ice sheets – something not possible before this system was developed.

EVOLVING COMPONENTS OF CYBERINFRASTRUCTURE INCLUDING CLOUD COMPUTING

In his 2007 article, Freeman stated that the definition of cyberinfrastructure will evolve over time (Freeman, 2007). Cloud computing can thus be thought of as a particular approach to computing infrastructure and as a component of cyberinfrastructure which includes computational resources and data storage resources. According to the National Institute of Standards and Technology (NIST),

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. (Mell & Grance, 2011)

Key in cloud computing are on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. It can be used as a solution to applications that are "hosted in the cloud" or integrated into cyberinfrastructure. Like a high performance computer, a cloud computing solution can be monolithic or integrated into a larger cyberinfrastructure facility. For example, the NSF recently funded a cloud system that will be integrated as part of XSEDE called Jetstream (Stewart et al., 2015). It can be used in isolation as a scientific cloud system or as part of a larger integrated cyberinfrastructure facility.

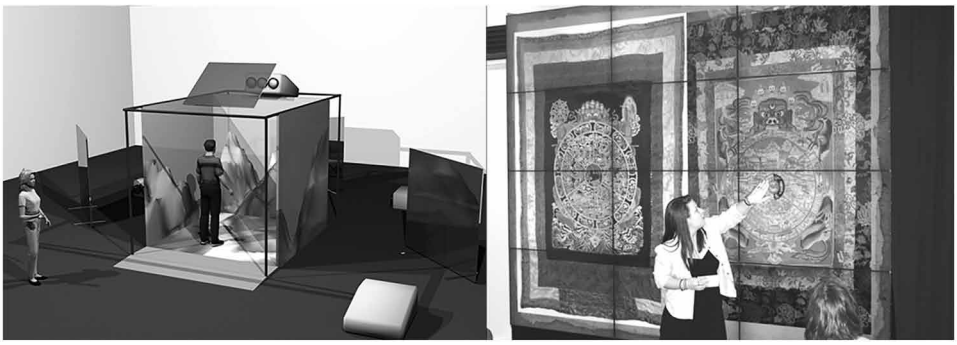
Data storage systems, advanced instruments and data repositories have also changed over time. Some recent changes in needs and data resources are described in the final report of the NSF Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization (NSF Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization, 2011) and the work

from Hey, Tansley, and Tolle (2009). The NSF has recently funded a new data-oriented storage and analytics facility called Wrangler to add new resources for big data to the US infrastructure funded by the NSF and supported by XSEDE.

Middleware has evolved significantly since the early days of cyberinfrastructure. Globus, one of the most widely used families of software in the world, now includes authentication, secure access capabilities, and data and metadata management tools (Foster, 2005; Globus Online, 2013). Other middleware includes workflow systems that coordinate the use of cyberinfrastructure and automate complex analyses; examples include Apache Airavata (Marru et al., 2011), Kepler (Ludäscher et al., 2006), and Pegasus (Deelman et al., 2005).

Visualization systems — hardware (display, visualization, and interaction) and software (applications, libraries, middleware, and data format standards) — have evolved dramatically since the inception of the term cyberinfrastructure. Visualization was one of the earliest cyberinfrastructure components to promote distributed applications and high levels of interoperability, largely because of the network of homogeneous CAVE Automatic Virtual Environments (CAVEs) and smaller devices using similar software launched in the last half of the 1990s (NCSA, 2001). Users at multiple sites could synchronously interact with the same data sets and observe remote participants via virtual avatars while communicating over IP-based audio and video channels. CAVEs and similar devices introduced new capabilities for understanding complex 3D and 4D data from other cyberinfrastructure resources. However, cost and scarcity, limited their impact on day-to-day scientific investigation. The 2000s saw affordable graphics cards, projectors, and high-definition, stereoscopic displays. Consumer-level technologies spurred a range of innovative systems for stereoscopic and ultra-resolution visualization (Sherman, O'Leary, Whiting, Grover, & Wernert, 2010), democratizing advanced visualization systems and techniques. Figure 3 shows a CAVE diagram and an ultra-high resolution tiled wall assembled from commodity HD televisions. Shown in the figure at right, scholars at Indiana University's Mathers Museum of World Cultures use an IQ-Wall to compare high-resolution images of textiles. This 3x4 wall is free-standing and was installed in a museum gallery in an afternoon. Such a display wall, which supports collaborative research, can now be created for a few tens of thousands of dollars, making them widely accessible in research environments.

Figure 3. At left is a CAVE, a room-scale visualization environment. At right is an ultra-high resolution tiled wall built in 2013 using commodity HDTV displays.
Source: © 2015, Trustees of Indiana University. Used with permission



FUTURE RESEARCH DIRECTIONS

Cloud Computing, Cyberinfrastructure, Exascale Computing, and the Economics of Computing

Cloud computing and more traditional HPCs (supercomputers) have complementary strengths and weaknesses. Cloud computing facilities may have internal networks of modest speed compared to supercomputers. On the other hand, cloud computing may be purchased in modest increments and are thus more accessible to a larger user community than supercomputers. Cloud computing is commonly used for "big data" applications, characterized by data volume, velocity, and variety (Laney, 2001). US President Obama's recent executive order (Obama, 2015) sets a new agenda for the creation of exascale computing facilities (capable of 10^{15} mathematical operations per second) while calling for joint development of exascale and big data/cloud computing facilities. Fox and collaborators (Fox, Qui, Kamburugamuve, Jha, & Luckow, 2015) depict many of the commonalities between cloud computing and HPC and propose an alignment and set of commonalities between HPC and big data stacks that can form a foundation for the sort of joint development of both approaches called for by President Obama in his recent executive order.

Cloud computing and HPC need not be an "either/or" choice. There are business cases for selecting cloud computing or local servers (Brumec & VrčEk, 2013; Marston, Li, Bandyopadhyay, Zhang, & Ghalsasi, 2011). However, choosing between cloud and HPC can be very complicated as cloud resources may not be able to support science applications that require low-latency internal networks or large amounts of memory as prerequisites, even if the cost of cloud computing seems lower than HPC per CPU hour. A locally owned HPC or HTC system can be acquired as a one-time cost where the capacity of the system limits the usage over time but remains useable for several years. In contrast, use of cloud computing may have a smaller cost for initial use but requires ongoing payments. This suggests tradeoffs in "locally owned" versus "cloud" that will suggest solutions strongly influenced by local conditions and financial systems at any given organization.

Cyberinfrastructure challenges include documenting return on investment and energy costs to operate at large scale. An analysis of return on investment in XSEDE is explored by Stewart and collaborators (Stewart et al., 2015). Energy costs and the economics of large-scale data centers help drive many activities into cloud computing. Security and data privacy are also concerns.

Science Gateways, Campus Bridging, and Cyberinfrastructure Ease of Use

For years, researchers accessed cyberinfrastructure exclusively through command-line interfaces. This sort of interface made it difficult to do long complex tasks and they were not particularly user friendly. Today, access to and the utility of cyberinfrastructure has been considerably expanded through the deployment of science gateways, use of cloud computing tools, and campus bridging. In particular, science gateways provide access to cyberinfrastructure to a broad set of users by employing graphical user interfaces and sophisticated tools for orchestrating computational workflows.

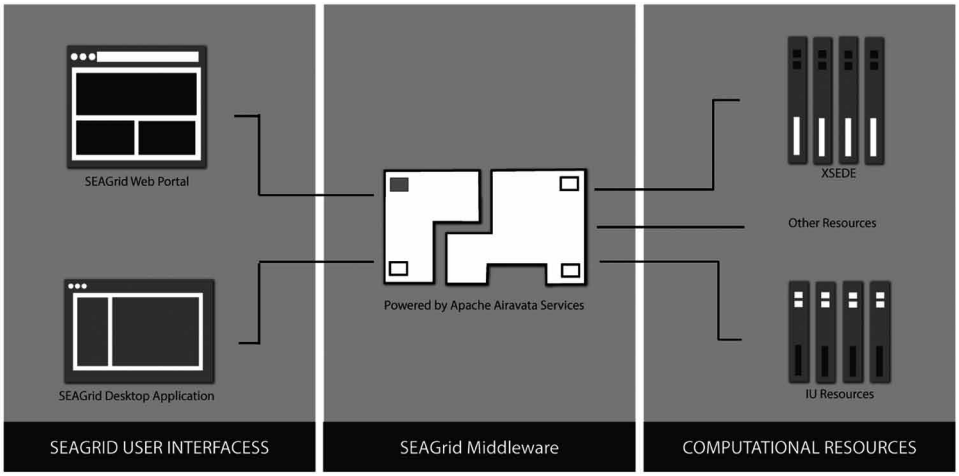
Science gateways make it possible to weave together a set of complicated tasks to achieve an overarching goal – like search for drug candidates or predict the path of a tornado. More formally, science gateways are defined as "a community-specific set of tools, applications, and data collections that are integrated together via a portal or a suite of applications" that can "support a variety of capabilities including workflows, visualization as well as resource discovery and job execution services" (Wilkins-Diehr, 2007). There are now dozens of science gateways in use or in development (Lawrence et al., 2015). For example, the CiPRES portal (Cyberinfrastructure for Phylogen-

tic Research) enabled thousands of researchers – including research students – to do sophisticated analyses of evolutionary histories from genetic information (CIPRES, 2016). Another important science gateway is SEAGrid – the Science and Engineering Applications gateway (SEAGrid, 2015). SEAGrid is geared toward chemical and mechanical analyses, and can, for example, be used to search for potential new drug candidates. The major components of SEAGrid (see Figure 4) exemplify a multi-tiered approach commonly used in science gateways.

Science gateways have also had a profound impact on citizen science—the public contribution to scientific discoveries (OpenScientist, 2011). Zooniverse is a web-based front end to several science gateways supporting citizen science. Dozens of projects use citizen science in weather, archaeology, biology, and medicine, where thousands of people help analyze research data – particularly image data – that might otherwise go unanalyzed for months or years into the future (Zooniverse, 2015).

Campus bridging approaches a different set of cyberinfrastructure issues. One of the significant challenges in cyberinfrastructure is integrating across scales of resources. Campus bridging focuses on integrating local, often modest scale cyberinfrastructure facilities with regional, national, and even international cyberinfrastructure resources. The goal of campus bridging is to "bridge" from local infrastructure to national resources in a way that makes national resources accessible (Hallock, Knepper, & Stewart, 2015; NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging, 2011). Some campus bridging issues are solvable simply with funding; for instance, networking becomes ever cheaper and it becomes ever more feasible to have good network connectivity from campus to national resources. Recent efforts have focused on technical interoperability among campus computing clusters and national level resources like XSEDE. The technical aspects of such in-teroperability include adding software to local institutional cyberinfrastructure systems that match those used on nationally-shared resources and as a result enabling training and educational materials developed for nationally shared systems to be used in support of smaller, local resources. Networking in support of bridging from campus to national resources has also been furthered by a network concept called the Science DMZ, which explicitly creates a portion of network "specifically engineered for science applications and does not include support for general-purpose use. By separating the high-performance science network

Figure 4. Science gateways such as SEAGrid uses a multi-tiered gateway architecture



(the Science DMZ) from the general-purpose network, each can be optimized without interfering with the other" (ESnet, 2016).

CONCLUSION

Cyberinfrastructure has evolved from supercomputer centers into an integrated and distributed suite of powerful and flexible resources that integrate supercomputers, data resources, visualization, and people in ways that go beyond the capabilities of any of the individual components of cyberinfrastructure. It has led to new products, medical treatments, and improved business processes that improve quality of life. In the long run we believe that cloud computing, high performance computing, and high throughput computing will be seen not as alternatives but as complementary tools used flexibly in response to the particular science and engineering needs and particular local conditions of researchers and organizations making use of cyberinfrastructure.

In summary, the future offers tremendous opportunities for science and society to use cyberinfrastructure to enable new discoveries and improve the quality of life of people everywhere as new tools for visualization, science gateways, campus bridging, citizen science, and cloud computing evolve and deliver new capabilities to the public and the scientific and technical communities worldwide.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under grants 0504075, 0451237, 0723054, 1062432, 0116050, 0521433, 0503697, 1053575, and ACI-1445604, and support provided by the Indiana University Pervasive Technology Institute. Any opinions, findings and conclusions or recommendations expressed herein are those of the authors and do not necessarily reflect the views of the supporting agencies. Robert Quick of IU created the OSG map in Figure 1. Editing by Greg Moore and Winona Snapp-Childs is gratefully acknowledged; any errors are the responsibility of the senior author.

REFERENCES

Atkins, D. E., Droegemeier, K. K., Feldman, S. I., Garcia-Molina, H., Klein, M. L., Messerschmitt, D. G., … Wright, M. H. (2003). *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Retrieved from http://www.nsf.gov/cise/sci/reports/atkins.pdf

Avery, P. (2007). Open science grid: Building and sustaining general cyberinfrastructure using a collaborative approach. *First Monday*, *12*(6). doi:10.5210/fm.v12i6.1866

Brumec, S., & Vrček, N. (2013). Cost effectiveness of commercial computing clouds. *Information Systems*, *38*(4), 495–508. doi:10.1016/j.is.2012.11.002

CIPRES. (2016). *CIPRES Cyberifrastructure for Phylogenetic Research*. Retrieved June 30, 2016, from http://www.phylo.org/sub_sections/portal/

Clarke, R., & Hunker, J. (1998). *Press Briefing by Richard Clarke, National Coordinator for Security, Infrastructure Protection, and Counter-Terrorism; and Jeffrey Hunker, Director of the Critical Infrastructure Assurance Office*. Retrieved from http://www.fas.org/irp/news/1998/05/980522-wh3.htm

Deelman, E., Singh, G., Su, M.-H., Blythe, J., Gil, Y., Kesselman, C., … Katz, D. S. (2005). Pegasus: A framework for mapping complex scientific workflows onto distributed systems. *Sci. Program., 13*(3), 219–237. Retrieved from http://dl.acm.org/citation.cfm?id=1239653

Dreher, P., Agarwala, V., Ahalt, S. C., Almes, G., Fratkin, S., Hauser, T., … Stewart, C. A. (2009). Developing a Coherent Cyberinfrastructure from Local Campuses to National Facilities: Challenges and Strategies. *EDUCAUSE*. Retrieved from http://www.educause.edu/Resources/DevelopingaCoherentCyberinfras/169441 or http://hdl.handle.net/2022/5122

ESnet. (2016). *Why Science DMZ*. Retrieved June 30, 2016, from https://fasterdata.es.net/science-dmz/motivation/

Foster, I. (2005). Globus Toolkit Version 4: Software for Service-Oriented Systems. In H. Jin, D. Reed, & W. Jiang (Eds.), Network and Parallel Computing (Vol. 3779, pp. 2–13). Springer Berlin Heidelberg. http://doi.org/doi:10.1007/11577188_2

Foster, I., & Kesselman, C. (1998). *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann.

Foster, I., & Kesselman, C. (2004). The Grid 2, Second Edition: Blueprint for a New Computing Infarstructure. Morgan Kauffman.

Fox, G. (2006). Collaboration and Community Grids. In *International Symposium on Collaborative Technologies and Systems* (pp. 419–428). IEEE Computer Society. http://doi.org/ doi:10.1109/CTS.2006.24

Fox, G. C., Qui, J., Kamburugamuve, S., Jha, S., & Luckow, A. (2015). HPC-ABDS High Performance Computing Enhanced Apache Big Data Stack. *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, 1057–1066. http://doi.org/doi:10.1109/CCGrid.2015.122

Freeman, P. A. (2007). Is Designing Cyberinfrastructure or, Even, Defining It Possible? *First Monday*, *12*(6). doi:10.5210/fm.v12i6.1900

Globus Online. (2013). *Globus Toolkit Homepage*. Retrieved from http://www.globus.org/toolkit/

Hallock, B., Knepper, R., & Stewart, C. A. (2015). *Workshop Report: Campus Bridging: Reducing Obstacles on the Path to Big Answers, 2015. IEEE Cluster 2015*. Retrieved from http://hdl.handle.net/2022/20538

Hayden, L., Fox, G., & Gogineni, P. (2007). Cyberinfrastructure for Remote Sensing of Ice Sheets. Madison, WI: TeraGrid. Retrieved from http://cerser.ecsu.edu/citeam/teragrid07.pdf

Johnston, W. E., Gannon, D., & Nitzberg, B. (1999). Grids as production computing environments: the engineering aspects of NASA's Information Power Grid. *High Performance Distributed Computing, 1999. Proceedings. The Eighth International Symposium on*, 197–204. http://doi.org/ doi:10.1109/HPDC.1999.805298

Knepper, R., Link, M. R., & Standish, M. (2015). Big Data on Ice: The Forward Observer System for In-flight Synthetic Aperture Radar Processing. *Procedia Computer Science, 51*, 1504–1513. Retrieved from http://hdl.handle.net/2022/20471

Korab, H., & Brown, M. D. (1995). Virtual Environments and Distributed Computing at SC'95: GII Testbed and HPC Challenge Applications on the I-WAY. In H. Korab & M. D. Brown (Eds.), *SUPERCOMPUTING '95*. New York, NY: ACM.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety. *META Delta*. META Group. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Litzkow, M., Livny, M., & Mutka, M. (1988). Condor - A Hunter of Idle Workstations. *8th International Conference of Distributed Computing Systems*, 104–111.

Livny, M., Avery, P., Pordes, R., Foster, I., & Lazzarini, A. (2006). *Sustaining and Extending the Open Science Grid: Science Innovation on a PetaScale Nationwide Facility*. National Science Foundation. Retrieved from http://nsf.gov/awardsearch/showAward.do?AwardNumber=0621704

Ludäscher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., & Zhao, Y. et al. (2006). Scientific workflow management and the Kepler system. *Concurrency and Computation*, *18*(10), 1039–1065. doi:10.1002/cpe.994

Marru, S., Gunathilake, L., Herath, C., Tangchaisin, P., Pierce, M., & Mattmann, C., … Weerawarana, S. (2011). Apache airavata: a framework for distributed applications and computational workflows. In *Proceedings of the 2011 ACM workshop on Gateway computing environments* (pp. 21–28). Seattle, WA: ACM. http://doi.org/doi:10.1145/2110486.2110490

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing - The business perspective. *Decision Support Systems*, *51*(1), 176–189. doi:10.1016/j.dss.2010.12.006

Mell, P., & Grance, T. (2011). *The NIST Definition of Cloud Computing (SP 800-145). Recommendations of the National Institute of Standards and Technology*. Retrieved from http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf

National Science Foundation. (2006). *Cyberinfrastructure: From Supercomputing to the TeraGrid*. Retrieved from http://www.nsf.gov/news/special_reports/cyber/fromsctotg.jsp

NCSA. (2001). *CAVERN Users Society*. Retrieved November 30, 2015, from http://cavernus.ncsa.illinois.edu/

NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging. (2011). *NSF Advisory Committee for Cyberinfrastructure Task Force on Campus Bridging Final Report*. Retrieved from http://www.nsf.gov/od/oci/taskforces/TaskForceReport_CampusBridging.pdf or http://pti.iu.edu/campusbridging/

NSF Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization. (2011). *NSF Advisory Committee for Cyberinfrastructure Task Force on Data and Visualization Final Report*. Retrieved from http://www.nsf.gov/od/oci/taskforces/TaskForceReport_Data.pdf

Obama, B. (2015). *Executive Order -- Creating a National Strategic Computing Initiative*. Retrieved from https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

Open Science Grid. (2015). *About the Open Science Grid*. Retrieved November 30, 2015, from http://www.opensciencegrid.org/

Sherman, W. R., O'Leary, P., Whiting, E. T., Grover, S., & Wernert, E. A. (2010). IQ-station: A low cost portable immersive environment. Lecture Notes in Computer Science, 6454, 361–372. http://doi.org/ doi:10.1007/978-3-642-17274-8_36

Stewart, C. A., Cockerill, T. M., Foster, I., Hancock, D., Merchant, N., & Skidmore, E., … Gaffney, N. (2015). Jetstream - A self-provisioned, scalable science and engineering cloud environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. http://doi.org/doi:10.1145/2792745.2792774

Stewart, C. A., Roskies, R., Knepper, R., Moore, R. L., Whitt, J., & 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* (pp. 23:1–23:8). New York, NY: ACM. http://doi.org/doi:10.1145/2792745.2792768

Stewart, C. A., Simms, S., Plale, B., Link, M., Hancock, D., & Fox, G. (2010). What is Cyberinfrastructure? SIGUCCS 2010. doi:10.1145/1878335.1878347

Tabor Griffin Communications. (1998). *High-Performance Computing Contributions to Society*. Tabor Griffin Communications. Retrieved from http://www.tgc.com/hpcbook

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: The Condor experience. *Concurrency and Computation*, *17*(2-4), 323–356. http://doi.org/doi doi:10.1002/cpe.938

The Apache Software Foundation. (2006). *Hadoop Distributed File System*. Retrieved May 6, 2014, from http://hadoop.apache.org/

Top500.Org. (2016). *The List*. Retrieved June 30, 2016, from https://www.top500.org/

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., … Wilkins-Diehr, N. (2014). XSEDE: Accelerating Scientific Discovery. *Comput. Sci. Eng., 16*, 62. doi:10.1109/MCSE.2014.80

U.S. Congress Office of Technology Assessment. (1993). *Advanced Network Technology--Background Paper*. Retrieved from https://www.princeton.edu/~ota/disk1/1993/9304/9304.PDF

XSEDE. (2013). *2013 Nobel Prize in Chemistry winners bring HPC to the lab*. Retrieved November 30, 2015, from https://www.xsede.org/2013-nobel-prize-in-chemistry

XSEDE. (2015). *Overview*. Retrieved November 30, 2015, from https://www.xsede.org/overview

KEY TERMS AND DEFINITIONS

Campus Bridging: The seamlessly integrated use of cyberinfrastructure operated with other local or remote cyberinfrastructure as if they were proximate to the user.

Citizen Science: The work of individuals or teams of amateur, non-professional, or volunteer scientists who conduct research, gather and analyze data, perform pattern recognition, and develop technology, often in support of professional scientists.

Cloud Computing: On-demand, affordable access to a distributed, shared pool of computing and storage resources, applications, and services usually via the Internet for a large number of users.

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

eScience: Computationally intensive science carried out through distributed global collaborations enabled by the Internet, involving access to large data collections, very large scale computing resources and high performance visualization.

High Performance Computing: Many tightly integrated computer processors that run very large scale computations and data analyses quickly where communication among the many processors is required.

High Throughput Computing: A computing paradigm that focuses on the efficient execution of a large number of loosely-coupled tasks

Science Gateways: Community-developed tools, applications, and data integrated via a portal or a suite of applications, usually in a graphical user interface, and customized to the needs of specific communities.