Jetstream: A Distributed Cloud Infrastructure for Under-resourced higher education communities

Jeremy Fischer
Indiana University
2709 E. Tenth Street
Bloomington, IN 47408
jeremy@iu.edu

Steven Tuecke
Computation Institute
University of Chicago
Argonne National Laboratory
Chicago, IL 60637
tuecke@ci.uchicago.edu

Ian Foster
Computation Institute
University of Chicago
Argonne National Laboratory
Chicago, IL 60637
foster@anl.gov

Craig A. Stewart
Indiana University
2709 E. Tenth Street
Bloomington, IN 47408
stewart@iu.edu

ABSTRACT

The US National Science Foundation (NSF) in 2015 awarded funding for a first-of-a-kind distributed cyberinfrastructure (DCI) system called Jetstream. Jetstream will be the NSF's first production cloud for general-purpose science and engineering research and education. Jetstream, scheduled for production in January 2016, will be based on the OpenStack cloud environment software with a menu-driven interface to make it easy for users to select a pre-composed Virtual Machine (VM) to perform a particular discipline-specific analysis. Jetstream will use the Atmosphere user interface developed as part of iPlant, providing a low barrier to use by practicing scientists, engineers, educators, and students, and Globus services from the University of Chicago for seamless integration into the national cyberinfrastructure fabric. The team implementing Jetstream has as their primary mission extending the reach of the NSF's eXtreme Digital (XD) program to researchers, educators, and research students who have not previously used NSF XD program resources, including those in communities and at institutions that traditionally lack significant cyberinfrastructure resources. We will, for example, use virtual Linux Desktops to deliver DCI capabilities supporting research and research education at small colleges and universities, including Historically Black Colleges and Universities (HBCUs), Minority Serving Institutions (MSIs), Tribal colleges, and higher education institutions in states designated by the NSF as eligible for funding via the Experimental Program to Stimulate Competitive Research (EPSCoR). Jetstream will be a novel distributed cyberinfrastructure, with production components in Indiana and Texas. In particular, Jetstream will deliver virtual Linux desktops to tablet devices and PDAs with reasonable responsiveness running over cellular networks. This paper will discuss design and application plans for Jetstream as a novel Distributed CyberInfrastructure system for research education.

Categories and Subject Descriptors

C.3 [**Computer Systems Operations**]: DISTRIBUTTED CYBERINFRASTRUCTURE; DCI; DCI OUTREACH; HCI; EASE OF USE; SPECIAL-PURPOSE AND APPLICATION-BASED SYSTEMS

K.4 [**Computing Milieux**]: Computing and Education

General Terms

Management, Documentation, Human Factors, Design, Standardization, Compatibility

Keywords

Cloud, extreme, digital, XSEDE, education, outreach, training, EOT, research, Jetstream, XD, Globus, Atmosphere, cyberinfrastructure

1. INTRODUCTION

US science and engineering has many important needs for a variety of forms of distributed cyberinfrastructure. While the NSF has provided many different types of distributed cyberinfrastructure, until 2015 it had never funded a production-quality cloud facility for support of general-purpose science and engineering research and research education. Jetstream is a new cloud system currently under construction, funded primarily through a grant award from the National Science Foundation [1, 2] to a partnership led by the Indiana University Pervasive Technology Institute (IUPTI) of Indiana University (NSF Award #ACI-1445604). Jetstream will be the first cloud facility funded by the NSF supporting all areas of science and engineering within NSF's scope. It will be a "science production" quality resource, rather than an experimental system supporting computer science research such as FutureGrid [3], Chameleon [4], or Cloudlab [5].

The NSF eXtreme Digital program, the successor to the TeraGrid [6], is an advanced, nationally distributed, open cyberinfrastructure comprised of various computational and scientific resources connected by high-bandwidth networks, integrated by coordinated policies and operations, and supported by computing and technology experts. XD directly enables and supports leading-edge scientific discovery and promotes science and technology education. [7].

The Extreme Science and Engineering Discovery Environment (XSEDE), perhaps the most visible part of the eXtreme Digital program of the NSF, is a project, an institution, and a set of services. XSEDE is a virtual organization that provides a distributed infrastructure, support services, and technical expertise, enabling researchers, engineers, and scholars to address challenging problems facing the scientific community, the nation, and the world. XSEDE supports a growing collection of advanced high-performance computing resources, high-end visualization resources, data analysis resources, as well as other resources and

[Event Title Goes Here Once Accepted]

[doi here]

services [8]. XSEDE manages the XD shared user and management services and is the final component of the XD program [7]

The NSF solicitation for the grant program that provided funds for Jetstream [1] expresses the NSF's view that needs of the US open science and research community have evolved more quickly than the diversity of resources available via the XD program and supported by XSEDE:

> *The current solicitation is intended to complement previous NSF investments in advanced computational infrastructure by exploring new and creative approaches to delivering computational resources to the scientific community. Consistent with the Advanced Computing Infrastructure: Vision and Strategic Plan (February 2012), the current solicitation is focused on expanding the use of high-end resources to a much larger and more diverse community. To quote from that strategic plan, the goal is to "... position and support the entire spectrum of NSF-funded communities ... and to promote a more comprehensive and balanced portfolio .... to support multidisciplinary computational and data-enabled science and engineering that in turn supports the entire scientific, engineering and educational community." Thus, while continuing to provide essential and needed resources to the more traditional users of HPC, this solicitation expands the horizon to include research communities that are not users of traditional HPC systems, but who would benefit from advanced computational capabilities at the national level.*

The primary goal set by IU and its partners in implementing Jetstream is to create a resource that expands the users of NSF eXtreme Digital (XD) program resources beyond the current community of users.

One particular area of focus within the Jetstream project is the development of a 21st century workforce that draws on the full richness of American society. Support for research and research education at colleges and universities that are under-resourced in terms of cyberinfrastructure. In some cases, tribal colleges or HBCUs operate with desktop workstations of a vintage appropriate for Microsoft Windows XP…and are still running Windows XP. The gap between needs generally for networking and actual installed capacity is described in Campus Bridging Data and Networking Issues Workshop Report [9]. Services to deliver cloud computing resources to under-resourced higher education communities is an issue that addresses cloud computing, research-as-a-service, and the development of a well educated and computationally literate 21st century workforce for the US from our own national community. With Jetstream implementation we will focus particularly on providing resources accessible by and useful to researchers and educators at under resourced and traditionally underserved institutions including HBCUs (Historically Black Colleges and Universities), MSIs (Minority Serving Institutions), Tribal colleges, and higher education institutions in EPSCoR States [10].

In the remainder of this paper, we explain how Jetstream will fulfill this goal by complementing existing NSF-funded cyberinfrastructure and in particular address the implementation of Jetstream in ways that will lead to adoption by researchers and research students at under-resourced postsecondary educational institutions throughout the US. Many of these implementation strategies will aid adoption within disciplines that have not historically made great use of NSF-funded supercomputers and cyberinfrastructure, but our emphasis in this paper is on researchers, educators, and students at Minority Serving Institutions (MSIs), including Historically Black Colleges and Universities (HBCUs) and Hispanic Serving Institutions (HSIs), as well as institutions in EPSCoR states [11]**.** In particular, then, this paper describes design principles and practices and novel application types that will be implemented through the NSF-funded Jetsream as a new and novel distributed cyberinfrastructure.

2. CYBERINFRASTRUCTURE AS A TECHNOLOGY ADOPTION CHALLENGE

Many factors must be considered when designing a resource that is accessible and usable by underserved communities. Fundamentally the decision as to whether or not to try to use some particular piece of technology is based on an evaluation of the value and cost of that technology. Our approach to implementation of Jetstream is based on current social science-based understandings of technology adoption [12] that suggests technology adoption is driven by: performance expectancy (perceived value), effort expectancy (perceived ease of use), social influence, and facilitating conditions (including knowledge of a technology and the belief that end users will find it accessible). In order to expand the US open science research community, the quantity and quality of research done in the US, and the annual rate of graduation of undergraduate and graduate degree holders that matriculate with a good understanding of computational science techniques, we as a nation must increase the adoption of cyberinfrastructure in post-secondary institutions not now offering curricula and graduate research programs based on computational science techniques.

Performance expectancy – perceived value of access to distributed cyberinfrastructure at post-secondary educational and research institutions with limited budgets – is perhaps the easiest problem to address in terms of encouraging adoption. There is widespread consensus that the US needs more people with a good background in computational science to help expand the US workforce [13, 14]. There is also clear educational research that says one of the best ways to interest students in science and engineering is to involve them in authentic scientific research as part of the educational process [14, 15]. Many publicly available "science-quality" data sets have yet to be fully mined for information and insights. Many highly qualified faculty researchers at small US colleges and universities, including at MSIs, could do important, original research analyzing data from such public data sources. Students could also participate in such research. However, in many cases the ability to obtain and work with this data is limited by local network capability, lack of local computing equipment, and lack of system administrator resources to support research computing. We have not been able to find a published survey of faculty at small schools about research, research education, and cyberinfrastructure. However in BOFs (Birds-of-a-Feather) sessions and in discussions with a large number of faculty at smaller schools, one message we hear consistently is (paraphrased) "if we had access to computing resources, we could do many interesting things, but for lack of that we don't even try to pursue computational science in our curriculum."

Effort expectancy - perceived ease of use - is a crucial factor in adoption of cyberinfrastructure and computational approaches to science and engineering by small and under-resourced postsecondary educational institutions. In many cases, local resources suitable for research and research education are simply not available, and nationally provisioned resources are not perceived to be accessible.

Perceived cost of use is a critical factor in perceived ease of use. Many smaller institutions have limited budgets and resources for workstations and IT staff. In developing our proposal to create Jetstream we visited with faculty and staff at a number of MSIs, including Tribal Colleges and MSIs. We talked with individuals at institutions where a years-old PC running the Windows XP operating system – an OS that was deprecated by Microsoft in 2014 after being in use for 12 years [16]. It is a sad and simple fact that there are many smaller institutions of higher education that do not have local desktop resources suitable for personal productivity or research use in 2015 – much less to such institutions have their own computing clusters or clouds.

Networking is another aspect of ease of use and accessibility. At larger educational institutions and government installations, we often take for granted our robust high-speed networks and ubiquitous wi-fi access. If it is impractical to do interactive computing or move data sets back and forth from campus to national facilities due to limitations of local network capabilities, then the ease of use of national resources and their accessibility will seem to be poor.

Perceived availability of the privilege to use national CI facilities is another aspect of perceived ease of use. XSEDE – the eXtreme Science and Engineering Discovery Environment – offers a very large cadre of computational and data-analysis tools available through peer review to the national research and education community. Use of XSEDE resources is done through allocations, and the allocations process is implemented through the XSEDE Resource Allocation System (XRAS) [17]. The existing XRAS process is geared toward making allocations of very large amounts of computing resources on very large leading-edge systems serving national communities. The process for making allocations when the monetary value of these allocations is commonly hundreds of thousands to millions of dollars per award seems daunting to a researcher who needs only a relatively modest – by national standards – amount of computing resource. The process of ensuring that multi-million dollar investments by the NSF are well utilized to support high quality, peer reviewed research can create a perception that these resources are in practice not accessible to faculty and students at many small and under-resourced institutions throughout the US.

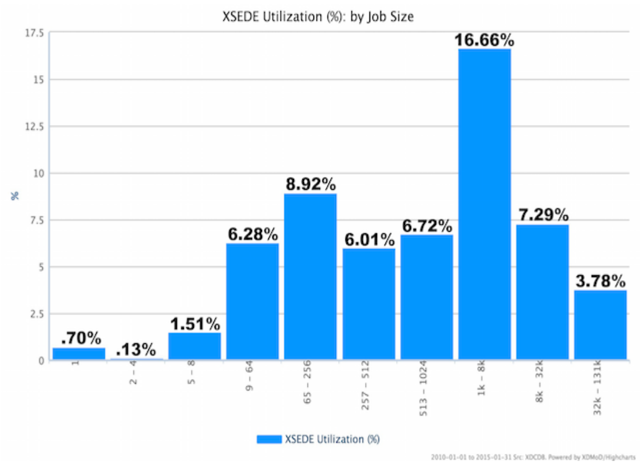
Social influence is an important aspect in encouraging adoption. Do people you know and trust make use of a particular resource? What has their experience been? When there are social influences promoting use of a particular technology choice adoption can be accelerated significantly.

Facilitating conditions- particularly training – is also a critical factor in encouraging adoption of technology. One can provide the best tools possible, but without intended users having the knowledge necessary to use them, they may go unused or, at the very least, underutilized.

3. JETSTREAM AS A COMPLEMENT TO EXISTING NSF-FUNDED RESOURCES

Jetstream is described in some detail online on the Jetstream Technical Backgrounder page [18] A brief description of the Jetstream hardware configuration from that page is that it "will consist of two homogenous clusters at Indiana University and TACC with a test environment at the University of Arizona. The system will provide over ½ a PetaFLOPS of computational capacity and 2 petabytes of block and object storage" [18]. The software side will have a web-based interface based on the Atmosphere cloud computing environment with the backend software running Openstack [18, 19].

Part of complementing the existing use of NSF XD program resources is understanding the current usage patterns of those resources. Figure 1 depicts the utilization of XSEDE resources from 2010-2015 based on data from the XD Metrics on Demand (XDMoD) tool [20].



2010-01-01 to 2015-01-31 Src: XDCDB. Powered by XDMoD/Highcharts

Figure 1. XSEDE utilization by job size/cores for 2010-2015

We are all familiar with the adage 'walk before you run.' That is as true of parallel computing and use of cyberinfrastructure as anything. Under NSF guidance, XSEDE has created a system of startup, education, and research allocations [21] as a way to east the learning curve to use of XD program resources. Education and startup allocations are granted to XSEDE members that apply with certain conditions and limitations on use for each. These two types of allocations may be requested at any time, though a PI may only have one startup allocation at a time. Research allocations have more rigid requirements and are subject to committee review but may provide considerably more use of resources [21]. From 2011-2014, 2,552 startup allocations were granted from XSEDE. During this same period, the XSEDE Resource Allocations Committee (XRAC) received 583 research allocations. Of these, 435 were approved (74.61%) and 148 (25.39%) were rejected [22].

Utilization of XD program resources is dominated by jobs using between 1,000 and 8,000 cores. Jobs using 1-4 cores constitute less than 1% of XSEDE systems utilization over the 5-year period from 2010-2015. One core through 8 core jobs combined only account for 2.34% of jobs total running on XD program resources.

Startup allocations are not intended to provide an ongoing supply of cycles, so users must either continue to apply for new and different startup allocations or write an XRAC proposal and hope to get an allocation through that method. As the XSEDE KB notes, "Although renewals for Startup allocations are permitted with appropriate justification, a PI with an expiring Startup allocation should consider requesting a Research allocation to continue work on XSEDE.” [23]

Jetstream seeks to solve this problem, in partnership with XSEDE, by creating a different type of resource, aimed not at solving the complicated “number-crunching” problems that typical high-performance computing (HPC) resources solve but rather allowing for on-demand resources aimed for smaller jobs, exploratory science to test theories without burning HPC allocation times, as well as creating virtual machine images for educating users about science and engineering resources and how to use them. Jetstream is configured in a way that emphasizes small jobs – for learning purposes and in order to support workloads that need a handful of processors “now” whenever now is. Jobs on Jetstream will run within VMs, where the VM sizing is designed to support workloads predominantly consistent of smaller jobs.

One approach to virtual machine configuration is to just have different sizes and allocations of resources (e.g. X cores, Y RAM, etc.) per VM. Another, perhaps more useful method, is to have VMs tailored to specific applications. Some of these use cases include ideas such as VMs tailored for use for the National Snow and Ice Data Center (NSIDC), making specific large-scale data sets readily available to users of that VM easily as well as common software for that analysis, including utilizing user-held licenses to make proprietary software available on those VMs. Another example might be a VM catering to GIS users - making software like ArcGIS available to those users and other software and data sets a GIS user may need.

The ultimate goal is to create VMs that look like a desktop environment that any researcher might use, making it easier to use than a command line, text only environment one might find on some high-performance computing resources, making the VMs specific where possible to remove distractions and to tailor the machine to specific uses, allowing the research/researcher to be more efficient while filling unmet needs in the scientific community. Instead of just making a resource primarily focused on the need for computational power or large memory, Jetstream VMs can focus on the smaller projects that need resources as well as the "long tail" of research, the long-term, on-going research that may follow initial discoveries or research going on in the wake of it for many years. This would help not only fill in the gaps of ongoing research but also allow the long-term research to continue after the initial research allocation on a larger XD resource is depleted.

4. ATMOSPHERE AND GLOBUS AS MECHANISMs FOR PROMOTING EASE OF USE

Atmosphere is a cloud service that lets you launch your own isolated virtual machine working environment. It’s a virtual desktop configured with specific software and possibly with specific data sets to give researchers a “tuned” environment to meet the needs of specific research. [19] Often, large shared clusters are not always able to provide customized execution environments for specific tasks and software tools. Atmosphere attempts to address these issues by providing preconfigured, domain-specific virtual instances of common software tools that are contained in other parts of iPlant's cyberinfrastructure. [24]

This allows Atmosphere users to quickly develop tools, workflows and algorithms, reducing the extensive time, resources, and overhead needed to prepare a computing resource to be used for specific analyses. Users can access portions of data and software required by a specific analysis from iPlant’s existing resources by including it into their virtual machine (VM) configuration. Users are also able to preserve the state of their VM configurations, saving the system state ad configuration, allowing this VM to be used by other researchers for extending or reproducing this specific research or workflow. [24]

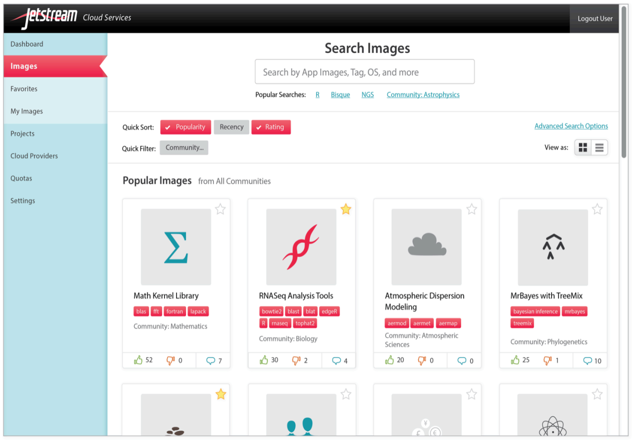


Figure 2 Shows what Atmosphere for Jetstream may look like for users. VM images would be tuned to users' needs.

We will apply four major methods to help Jetstream achieve the goal of providing resources to the underserved. First, Jetstream will engage the services of the Cornell Virtual Workshop program to create training materials for researchers that want to use the system. In conjunction with the standard user guides and Knowledge Base (KB) entries generated for XSEDE/XD resources, this will form the first pillar of education for those wishing to use Jetstream.

Second, we will utilize the services of the XSEDE Campus Champions program. The Campus Champions serve as a local resource on a campus to researchers. Campus Champions are “a local source of knowledge about high-performance and high-throughput computing and other digital services, opportunities and resources.” [25]

Third, the Cornell University Center for Advanced Computing will create virtual workshops focusing on Jetstream. Cornell has developed a number of workshops on high-performance computing topics that include system architecture, parallel programming, data management and other topics [26]. The educational benefit to the XSEDE community is broad and gives researchers a self-driven method to further their knowledge or refresh themselves on topics that will help them use XD resources more efficiently and to their maximum benefit.

The fourth, and arguably most important, method will be to make use of the existing XSEDE Training, Education, and Outreach Services (TEOS). Through campus visits, online and in person classes, outreach events and conferences, and campus bridging, the TEOS service “works with campuses and organizations to help instill digital services into the practices of faculty, staff and students.” [27] This mechanism is fully developed and in place to help users of XD resources and services increase their knowledge of not only using these resources but also using them most efficiently as well as teaching new software and programming languages and techniques that they may apply to research.

We leverage Globus services [28, 29] to facilitate integration of Jetstream into the national cyberinfrastructure. In particular, we use Globus transfer services for rapid, efficient, and secure file movement synchronization between Jetstream, campus systems, and other national cyberinfrastructure; Globus sharing for authorizing remote access to data stored on Jetstream; and Globus identity and group management services [30] to permit access to Jetstream with campus credentials and to manage groups used for access control.

Globus services, accessible at www.globus.org, apply software-as-a-service (SaaS) methods to deliver powerful research data management capabilities to many users at low cost. As with commercial SaaS products, users need no client software to access Globus services. Instead, users employ an intuitive web interface (or, if they prefer, REST or command line interfaces) to request file transfer, synchronization, or sharing operations. Once a request is made, Globus handles all required authentication, configures file transfers for high performance, and performs integrity checks (and, if requested, encryption). Globus also monitors the transfer over time, retrying if errors are detected. Users find this extreme ease of use and “fire and forget” model extremely useful, which probably accounts for the more than 23,000 registered users as of early 2015 and the fact that these users have so far used Globus services to transfer more than 80 PB in 9B files.

Globus services facilitates bridging from campus systems to national CI capabilities, including Jetstream [31]. A Globus endpoint (server software) is easily deployed on a campus system and integrated with campus authentication, so that users can employ campus credentials to access data via Globus,

5. JETSTREAM VIRTUAL DESKTOPS TO INCREASE EFFORT EFFICIENCY – PERCEIVED AND REAL EASE OF USE

In designing Jetstream we spent a considerable amount of time with the question: ‘How can we make a nationally-allocated XD program resource accessible, in practical terms, to a researcher, educator, or student at an under resourced college or university, where the local desktop systems may be as outdated as obsolete PCs running Windows XP, extremely limited aggregate network bandwidth, and there are no local CI resources for research or research education?’

Our answer: deliver a virtual Linux desktop from Jetstream to users on under resourced campuses over a wireless network to an inexpensive a terminal device as possible, in a way that allows the user to treat the equipment in front of them as an I/O device. All processing will take place on Jetstream or on other XD program resources via Science Gateways [32]. Assume that data sources will be primarily from publically accessible data sources, so movement of data will be from data source to Jetstream or other XD program resource via high-speed national networks

Our conceptual solution then was to create a VM image that would deliver a remote desktop to a user endpoint device in a way that provides reasonable performance using the least expensive possible terminal equipment. Remote desktop delivery corresponds to a campus bridging use case already described and analyzed in some detail [33, 34].

The first author did initial testing of remote Linux desktop systems from the Indiana University Quarry cluster [35] using NoMachine NX client to deliver a virtual Linux Desktop operating within a VM on the Quarry cluster. We performed tests of inexpensive (< $250) end user devices and use of the Virtual Linux Desktop over wireless networks. The purpose of these tests was to determine whether using a low power mobile device like a tablet in a variety of network environments was a viable option.

Table 1 describes the configurations and prices of the devices we tested. There are two sets of configurations. The first two are laptop options; the first being an older Intel Core 2 Duo CPU model that meets the requirements for using VNC and is one of the more affordable new laptop options available still. The second is a newer budget model utilizing an Intel Mobile Celeron (Bay Trail-M) processor. This is new from Dell and does not have any educational discount applied. The second set of devices is Android tablets. These were chosen for their affordability/value. The first model specified is 3G capable, removing the networking restriction as a barrier to using Jetstream. The second is a newer Dell Android tablet. These devices represent the minimum resources needed to access and use Jetstream. The starting price of $128 becomes affordable for most educational organizations. In addition, as prices continue to drop, other devices that will be compatible with Jetstream, including 3G capable devices will drop under the $100 mark if they have not already by the time this paper is published.

Table 1. Specifications and hardware prices for minimal systems to run remote virtual desktops from Jetstream via VNC. Price shows list retail price as of January 2015.

OS	Machine specifications	Source	Price
Windows or Linux	Dell Latitude E5500 Notebook. Intel Core 2 Duo P8700 2.53GHz, 2GB DDR2 Memory, 160GB HDD, DVDRW, 15” Display	tigerdirect.com	$169
Windows or Linux	Dell Inspiron 15 Intel® Celeron® processor N2830 (1M Cache, 2.41GHz), 4GB RAM, 500GB HDD, 15" Display, Windows 8.1	dell.com	$249
Android	Supersonic SC-77TV 7IN Android 4.2 Touchscreen tablet, dual core CPU, 3G	tigerdirect.com	$128
Android	Dell Venue 8 - 8" Android 4.4 Wi-Fi/Bluetooth, 16GB RAM, Internet Tablet, 2.1GHz Atom	dell.com	$149

Remote desktop functionality as a primary means of interface is valuable if the remote desktops are usable to the user in terms of performance and responsiveness versus perceived lag. To benchmark performance, we tested the time to load remote desktop images on an Apple iPad over a number of different networks at a number of different locations. Both cellular data based connections and wifi based connections were tested as part of this benchmarking.

We performed wifi tests in campus buildings and commercial spaces such as coffee shops. 4G LTE (Long Term Evolution) is the best standard service for devices connected to a cellular network. 3G is the previous, significantly slower standard. Edge networks (“Enhanced Data rates for GSM Evolution”) are the slowest cellular connection. The goal for this benchmarking was to perform multiple tests in wi-fi as well as each type of cellular data network: 4G LTE, 3G, and Edge.

Table 2 shows the details for three geographic areas (including one area in an EPSCoR state). We found the remote desktop loads reliably in 5 seconds or less over LTE or wifi connections, and under 20 seconds on 3G and Edge networks. In general, a wi-fi connection and 4G LTE connections were not noticeably different in general performance. The 3G and Edge launches and subsequent use showed the networks to be quick enough that the remote desktop functionality we propose to deliver will be useful to the intended users.

Table 2. Virtual desktop instantiation benchmarks – means ± 95% confidence intervals [1]

Location	Wi-Fi	4G LTE	3G	Edge
Bloomington, IN	4.11 ± 0.310 s	4.82 ± 0.199 s	13.71 ± 1.204 s	19.74 ± 1.632 s
Somerset/Lexington, KY	4.38 ± 0.621 s	9.78 ± 7.378 s	10.457 ± 0.698 s	16.07 ± 2.367 s
San Antonio, TX	5.85 ± 0.686 s	13.51 ±1.944 s	17.46 ± 0.818 s	N/A

This real world connection speed research combined with finding value solutions for hardware help show the viability of Jetstream as a resource accessible to a very broad audience. Many current resources are accessible via terminal/command line shell and thus can be utilized from desktop and laptop computers, tablets, and even modern smartphones. Jetstream will add a more user-friendly Linux desktop environment, that many users will most likely be more familiar with. This gives a broader appeal and a lower barrier to entry in terms of money for the user client, network speed necessary, and last, but certainly not least, with a lower level of specialized expertise needed to use research resources.

4. SOCIAL INFLUENCES AND THE JETSTREAM TEAM

The Jetstream team was brought together specifically to collect the technical challenges in deploying, operating, and supporting a production science cloud and also to create a partnership that would create effective social influence to attract and engage researchers and students at MSIs and at institutions in EPSCoR states. At the risk of employing stereotypes, it is difficult for a white, middle-aged computer science faculty member from a well funded school – the product of an American middle-class or better upbringing - to be seen as an effective role model by non-Caucasian faculty and students at small and often budget-limited MSIs and postsecondary institutions in EPSCoR states.

The Jetstream team includes investigators and researchers from a number of computing and research institutions. Indiana University leads the project. The Texas Advanced Computing Center (TACC), Computation Institute at the University of Chicago, and University of Arizona are partners in Jetstream. Other collaborators include the University of Texas San Antonio, Johns Hopkins University, Penn State University, Cornell University, University of Arkansas Pine Bluff, The National Snow and Ice Data Center (NSIDC), the Odum Institute of North Carolina, and the University of Hawaii. In addition, Dell and Rackspace are commercial partners to the Jetstream project [36].

The team includes representatives of two EPSCoR states, Hawaii and Arkansas. Hawaii is an unfunded collaborator, focused on oceanographic research and outreach to EPSCoR states. Dr. Gwen Jacobs of University of Hawaii is the inaugural chair of the Jetstream Stakeholder Advisory Board. The University of Arkansas at Pine Bluff is an HBCU with a diverse student population, founded with a land grant in 1890 [37] Dr. Jessie Walker of UAPB will take a lead role in Jetstream in cybersecurity research and teaching – his personal area of expertise – as well as outreach to MSI. The University of Texas at San Antonio is an unfunded collaborator in the initial construction of Jetstream, but will have funding during the Operations and Management phase beginning in 2016. UTSA is involved in integration of OpenStack features and OpenStack support. The personnel funded at UTSA will work under the direction of Dr. Paul Rad and work will be done primarily by Hispanic Ph.D. students working in his Cloud and Big Data Lab [38] The Jetstream team then includes a set of national leaders who are all expert in a particular area of cyberinfrastructure or science using cyberinfrastructure, and who represent sub communities within the US higher education community that represent institutions in states that take in small amounts of NSF funding and / or who represent significant populations of people defined by the NSF as members of traditionally underserved groups.

This purposeful approach to constructing the Jetstream team will enable us to leverage social influence within communities in under-resourced institutions in the US to encourage and accelerate adoption of Jetstream.

5. EXTENSIBILITY AND COMMERCIAL OPTIONS

VMs can either be "spun up" from a customizable pre-populated library or launched from an archived image, recreating a consistent user experience on other cloud systems should be possible. This would allow multiple researchers to work independently to attempt to reach similar results. Our initial results in migration of VM images from the Atmosphere interface to Amazon Web Services (AWS) show that many VMs created for Atmosphere simply run successfully on AWS. In order to make it possible for researchers and students to store, retrieve, and VMs (either unpopulated with data, or populated with scripts, data, and output files for reproducibility) Jetstream will allow users to preserve VMs with Digital Object Identifiers (DOIs). This will enable sharing of results, reproducibility of analyses, and new analyses of published research data. IU will store and make available these VMs and associated data in the IUScholarWorks archive for the foreseeable future, where they will be easily discoverable via Globus Publication services [1].

AWS is a commercial service, which on the one hand limits access but on the other hand promotes extensibility. This means that tools developed within Jetstream can be deployed – funding available – beyond the boundaries of the Jetstream hardware funded by the NSF.

We will also look at commercial offerings for delivery of Linux Virtual Desktops. We are particularly interested in the possibility of using Citrix virtualization tools delivered from Windows-based VMs as a way to promote extensibility of access to Linux Virtual Desktops. Citrix now offers Linux virtual applications and desktops delivered from Citrix XenApp and XenDesktop. The addition of the Linux Virtual Delivery Agent (VDA) allows adminstrators to easily integrate Linux VMs into their existing Windows infrastructure. One potential advantage here is that Citrix is a mature product that has been delivering virtual machines and desktops for a number of years. [39] Another advantage might be allowing the use of Windows desktops from the start rather than as a potential feature in the future.

6. CONCLUSIONS

Jetstream will be a first-of-a-kind implementation of a production-quality cloud for science and engineering research and research education across all disciplines supported by NSF-funded cyberinfrastructure. With the implementation of Jetstream we hope to dramatically expand the current base of users of the NSF XD program. In particular we plan to drastically decrease barriers to adoption of advanced distributed cyberinfrastructure at small and under-resourced colleges and universities, particularly MSIs, HBCUs, HSIs, Tribal colleges, and institutions in EPSCoR states. We are purposefully employing a set of strategies to support and encourage adoption that are based on current sociological understanding of technology adoption. We will also reduce perceived barriers to adoption by creating processes for requesting and using Jetstream that are straightforward for people without prior experience using NSF-funded cyberinfrastructure. Creating an intuitive user interface is a key focus for involvement of the Atmosphere interface and use of Globus, and subsequently the overall Jetstream strategy, helping create a shallower learning curve to using this system.

We will encourage adoption of Jetstream through the creation of training materials and make use of social influences to increase adoption among communities that do not traditionally make extensive use of NSF-funded cyberinfrastructure. By working with partners that have significant experience in HPC at extreme scales and those that have experience in virtualization while utilizing our own strengths in HPC, High Throughput Computing (HTC), storage, and resource availability, we anticipate creating a robust service that will be accessible to all. In working with partners from the underserved communities from the beginning of Jetstream, we aim to position the operational system to meet the needs of those currently not well represented in the research community today. As Jetstream becomes the first production science and engineering cloud for the National Science Foundation, we believe it will create a new model for future distributed cloud infrastructures supporting academic research and research education.

7. ACKNOWLEDGMENTS

This document was developed with support from National Science Foundation (NSF) grant ACI-1445604. Funding through NSF grant OCI-1053575 for campus bridging activities has helped inform our understanding of distributed CI needs at under-resourced campuses. This research was also supported by the Indiana University Pervasive Technology Institute and major funding for IUPUI from the Lilly Endowment, Inc. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF or other funding agencies.

Atmosphere is developed by The iPlant Collaborative under grant NSF #DBI – 1265383. Atmosphere documentation and suggestions for inclusion of portions of that documentation were made by Nirav Merchant (nirav@email.arizona.edu).

The term “Stakeholder Advisory Board” was suggested by Dr. Miron Livny of the University of Wisconsin and we gratefully acknowledge this insight and suggestion.

Concepts and material regarding Citrix were provided by Stephanie Cox (coxst@iu.edu), and her review of the manuscript is greatly appreciated.

Virtual machine configuration and maintenance during the testing phases was done by Joe Rinkovsky (jrinkovs@iu.edu) of the IU Research Technologies High Performance Systems group.

Technical review of the Jetstream topics and paper in general was done by David Hancock (dyhancoc@iu.edu), manager of the IU Research Technologies High Performance Systems group. His review was instrumental to the later revisions of this paper

8. REFERENCES

1. Stewart, C.A., Merchant, N., Foster, I.,Taylor, J.,Vaughn, M., *High Performance Computing System Acquisition: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment*. 2014, National Science Foundation Directorate for Computer and Information Science and Engineering - Advanced Cyberinfrastructure Division Award #1445604.
2. National Science Foundation. *High Performance Computing System Acquisition: Jetstream - A Self-Provisioned, Scalable Science and Engineering Cloud Environment*. 2014 [cited 2015 February 17]; Available from: http://www.nsf.gov/awardsearch/showAward?AWD_ID=1445604.
3. Stewart, C.A. *FutureGrid: an experimental, high-performance grid testbed*. National Center for Supercomputer Applications 2009 21 Oct 2009; Available from: http://hdl.handle.net/2022/13899.
4. *Chamelon: A configurable experimental environment for large-scale cloud research*. Available from: https://www.chameleoncloud.org/.
5. *Cloudlab*. [cited 2015 February 18]; Available from: http://www.cloudlab.us/.
6. Catlett, C., et al., *TeraGrid: Analysis of Organization, System Architecture, and Middleware Enabling New Types of Applications.*, in *Advances in Parallel Computing Volume 16, 2008: High Performance Computing and Grids in Action*, L. Grandinetti, Editor. 2008, IOS Press: Amsterdam.
7. National Science Foundation. *Directorate for Computer and Information Science and Engineering (CISE) Budget*. 2014 [cited 2015 February 18]; Available from: http://www.nsf.gov/about/budget/fy2014/pdf/18_fy2014.pdf.
8. John Towns, T.C., Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, Nancy Wilkins-Diehr, *XSEDE: Accelerating Scientific Discovery*. Computing in Science & Engineering. **16**(5): p. 62-74.
9. Indiana University. *Campus Bridging Data and Networking Issues Workshop Report*. 2010; Available from:

http://pti.iu.edu/workshops/campusbridging/docs/cb_datanet_online_full.pdf.

10. Stewart, C.A., *Big Data: Where can EPSCoR states use big data and what tools do EPSCoR states need to thrive?*, in *Presentation before the EPSCoR / IDEA Board of Directors*. 2014: Arlington, VA http://hdl.handle.net/2022/19210.

11. National Science Foundation. *About EPSCoR*. 24 Jan 2012]; Available from: http://www.nsf.gov/od/oia/programs/epscor/about.jsp.

12. Venkatesh, V., Morris, M.G., Davis, F.D., Davis, G.B, *User Acceptance of Information Technology: Toward a Unified View*. MIS Quarterly, 2003. **27**(3): p. 425-478.

13. Adkins, R.C. *America Desperately Needs More STEM Students. Here's How to Get Them*. 2012 [cited 2015 February 19]; Available from: http://www.forbes.com/sites/forbesleadershipforum/2012/07/09/america-desperately-needs-more-stem-students-heres-how-to-get-them/.

14. *Analysis: The exploding demand for computer science education, and why America needs to keep up*. [cited 2015 February 19]; Available from: http://www.geekwire.com/2014/analysis-examining-computer-science-education-explosion/.

15. Committee on Highly Successful Schools or Programs in K-12 STEM Education, National Research Council, *Successful K-12 STEM Education: Identifying Effective Approaches in Science, Technology, Engineering, and Mathematics*. 2011: The National Academies Press.

16. Microsoft Corp. *What does it mean if Windows isn't supported?* [cited 2015 February 17]; Available from: http://windows.microsoft.com/en-us/windows/help/what-does-end-of-support-mean.

17. XSEDE. *Allocations*. [cited 2014 7 May]; Available from: http://www.xsede.org/allocations.

18. Indiana University. *Jetstream Technical Backgrounder*. 2014 [cited 2015 February 18]; Available from: http://pti.iu.edu/jetstream/leaders.php.

19. iPlant Collaborative. *Atmosphere*. [cited 2015 February 18]; Available from: http://www.iplantcollaborative.org/ci/atmosphere.

20. XSEDE. *XDMoD: Comprehensive HPC System Management Tool*. [cited 2015 February 18]; Available from: https://xdmod.ccr.buffalo.edu/.

21. XSEDE. *Allocations Overview*. [cited 2015 January 13]; Available from: https://portal.xsede.org/allocations-overview.

22. Hackworth, K., *Finding data (Personal communiction + Spreadsheet)*, J.L. Fischer, Editor. 2014.

23. XSEDE. *What types of XSEDE allocations are available?* [cited 2015 January 12]; Available from: https://portal.xsede.org/knowledge-base/-/kb/document/anql.

24. Skidmore, E., et al., *iPlant atmosphere: a gateway to cloud infrastructure for the plant sciences*, in *Proceedings of the 2011 ACM workshop on Gateway computing environments*. 2011, ACM: Seattle, Washington, USA. p. 59-64.

25. XSEDE. *XSEDE Campus Champions*. 20 Feb 2012]; Available from: https://www.xsede.org/campus-champions.

26. Cornell University Center for Advanced Computing. *Cornell Virtual Workshop*. [cited 2015 February 19]; Available from: https://www.cac.cornell.edu/vw/.

27. XSEDE. *Training, Education and Outreach Services (TEOS)*. [cited 2015 February 18]; Available from: https://www.xsede.org/education-and-outreach.

28. Allen, B., Bresnahan, John, Childers, Lisa, Foster, Ian, Kandaswamy, Gopi, Kettimuthu, Raj, Kordas, Jack, Link, Mike, Martin, Stuart, Pickett, Karl, Tuecke, Steven., *Software as a Service for Data Scientists*. Communications of the ACM, 2012. **55**(2): p. 81-88.

29. Foster, I., *Globus Online: Accelerating and democratizing science through cloud-based services*. IEEE Internet Computing, 2011(May/June): p. 70-73.

30. Chard, K., et al. *Globus Nexus: Research Identity, Profile, and Group Management as a Service*. in *e-Science (e-Science), 2014 IEEE 10th International Conference on*. 2014.

31. Foster, I., et al. *Campus Bridging Made Easy via Globus Service*. in *XSEDE '12 Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the campus and beyond*. 2012. Chicago, IL: ACM.

32. Stewart, C.A., Knepper, R. D., Link, M. R., Pierce, M., Wernert, E. A.,. Wilkins-Diehr, N. , *Cyberinfrastructure, Science Gateways, Campus Bridging, and Cloud Computing*, in *Encyclopedia of Information Science and Technology, Third Edition*. 2014, http://www.igi-global.com: Hershey. PA.

33. Stewart, C.A., et al., *Campus Bridging Use Case Quality Attribute Scenarios*. 2012: http://hdl.handle.net/2022/14476.

34. Stewart, C.A., Knepper, Richard, Grimshaw, Andrew, Foster, Ian, Bachmann, Felix, Lifka, David, Riedel, Morris, Tueke, Steven., *XSEDE Campus Bridging Use Cases*. 2012. p. 22.

35. Indiana University. *Quarry at Indiana University*. [cited 2015 February 18]; Available from: https://kb.iu.edu/d/avkx.

36. Indiana University. *Jetstream Partners and Collaborators*. 2014 [cited 2015 February 18]; Available from: http://pti.iu.edu/jetstream/partners-collaborators.php.

37. *University of Arkansas at Pine Bluff*. [cited 2015 February 2015]; Available from: http://www.uapb.edu/.

38. Fish, C. *UTSA opens Cloud and Big Data Laboratory to support computing research, training*. 2013 [cited 2015 February 17]; Available from: http://www.utsa.edu/today/2013/11/clouddata.html.

39. Citrix Systems. *Citrix Offers Technology Preview of Linux Virtual Apps and Desktops Delivered from XenApp and XenDesktop*. [cited 2015 February 18]; Available from: http://www.citrix.com/news/announcements/aug-2014/citrix-offers-technology-preview-of-linux-virtual-apps-and-deskt.html.