Jetstream: A self-provisioned, scalable science and engineering cloud environment

Craig A. Stewart[a]
stewart@iu.edu

Timothy M. Cockerill[b]
cockerill@tacc.utexas.edu

Ian Foster[c]
foster@mcs.anl.gov

David Hancock[d]
dyhancoc@iu.edu

Nirav Merchant[e]
nirav@email.arizona.edu

Edwin Skidmore[e]
edwin@iplantcollaborative.org

Daniel Stanzione[b]
dan@tacc.utexas.edu

James Taylor[f]
james@taylorlab.org

Steven Tuecke[c]
tuecke@ci.uchicago.edu

George Turner[d]
turnerg@iu.edu

Matthew Vaughn[b]
vaughn@tacc.utexas.edu

Niall I. Gaffney[b]
ngaffney@tacc.utexas.edu

[a]IU Pervasive Technology Institute and IU School of Informatics and Computing

[b]Texas Advanced Computing Center
University of Texas at Austin
Road A, Austin, TX 78758

[c]Computation Institute
University of Chicago
5735 S. Ellis Ave Chicago IL 60637

[d]IU Pervasive Technology Institute
2709 E. Tenth Street
Bloomington, IN 47408-2671

[e]University of Arizona
1401 East University Boulevard
Tucson, AZ 85721

[f]Johns Hopkins University
Department of Biology
3400 N Charles St.
Baltimore MD, 21218

ABSTRACT

Jetstream will be the first production cloud resource supporting general science and engineering research within the XD ecosystem. In this report we describe the motivation for proposing Jetstream, the configuration of the Jetstream system as funded by the NSF, the team that is implementing Jetstream, and the communities we expect to use this new system. Our hope and plan is that Jetstream, which will become available for production use in 2016, will aid thousands of researchers who need modest amounts of computing power interactively. The implementation of Jetstream should increase the size and disciplinary diversity of the US research community that makes use of the resources of the XD ecosystem.

Categories and Subject Descriptors

B.8 [**Performance and reliability**]; C.2.4 [**Distributed Systems**]; D.2.13 [**Reusable Software**]; D.4.2 [**Storage Management**]

XSEDE'15, July 26 - 30, 2015, St. Louis, MO, USA.

ACM 978-1-4503-3720-5/15/07 …$15.00
DOI: http://dx.doi.org/10.1145/2792745.2792774

General Terms

Management, Measurement, Documentation, Performance, Design, Reliability, Human Factors, Standardization.

Keywords

Cloud computing, atmosphere, long tail of science, big data.

1. INTRODUCTION

The national cyberinfrastructure (CI) funded by the National Science Foundation (NSF) successfully serves thousands of researchers who are advancing critical areas of science and engineering. However, there are thousands of other NSF-supported researchers who are not using NSF-funded CI facilities, including the resources of the NSF-funded XD (eXtreme Digital) ecosystem—that is, the Extreme Science and Engineering Discovery Environment (XSEDE) [1, 2] and the computational, data analysis, visualization, and consulting services delivered by Service Providers that are part of the XSEDE Federation [3]. The XD ecosystem has traditionally focused on delivering high performance computing (HPC) resources, recently adding focus on high-throughput computing (HTC). In many cases the researchers who are not making use of XD ecosystem resources are those with computing and data analysis needs that do not fit the traditional HPC or HTC computing models for which most XD ecosystem resources have traditionally been optimized.

Recognizing the need for more diversity in CI resources in the XD ecosystem, in 2013 the NSF funded the Comet [4] system and the Wrangler data storage and data analytics system [5]. Comet is a large, heterogeneous cluster that supports virtual clustering to enable research in the "long tail of science." Wrangler will create unprecedented data manipulation and analysis capabilities within NSF-funded CI. The need for more diversity in future CI resources has recently been emphasized in the report by the Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020 [6]. In late 2014 the NSF continued enhancing diversity of the CI resources within the XSEDE ecosystem by funding the Bridges system, to be operated by the Pittsburgh Supercomputing Center, and Jetstream, to be operated by a multi-institution collaboration led by Indiana University (IU). Bridges will be a very useful heterogeneous data analytics system and is described in detail elsewhere [7,8].

In this paper we describe the motivation for proposing Jetstream, the configuration of the Jetstream system as funded by the NSF, the team that is implementing Jetstream, and the communities that we expect to make use of this new system. Jetstream will be NSF's first cloud computing system for use by all disciplines supported by the XD ecosystem [9]. Jetstream is intended to be far reaching in scope, focusing on interactive delivery of resources from a "production-quality" cloud resource. Implementation of Jetstream is led by the Indiana University Pervasive Technology Institute (PTI) [10] with a large group of institutional collaborators involved in the construction of this resource: University of Texas at Austin's Texas Advanced Computing Center (TACC), University of Chicago, University of Arizona, and Johns Hopkins University. During the operations and management phase slated to begin when Jetstream goes into production use in January 2016, the list of partners and collaborators will expand to include Cornell University, the National Snow and Ice Data Center, the Odum Institute at the University of North Carolina, the PTI-led National Center for Genome Analysis Support (NCGAS), Pennsylvania State University, University of Arkansas at Pine Bluff, University of Hawaii, and University of Texas at San Antonio.

2. JETSTREAM'S FUNCTIONS

Jetstream will be a configurable, large-scale cloud computing resource that leverages both on-demand and persistent virtual machine technology to support a wide array of software environments and services. As a fully configurable cloud resource, Jetstream bridges the obvious major gap in the current XD ecosystem, which has machines targeted at large-scale, high performance computing, high-memory, large-data, high-throughput, and visualization resources, but no general purpose cloud resources. Jetstream will:

- provide self-serve academic cloud services, enabling researchers or students to select a virtual machine (VM) image from a published library, or alternatively to create or customize their own virtual environment for discipline- or task-specific personalized research computing.
- host persistent VMs to support delivery of science gateways. Galaxy will be one of the initial supported science gateways.
- support data movement with Globus Connect and authentication via Globus Auth.
- enable new modes of sharing computations, data, and reproducibility. Jetstream will support publication and sharing of VMs via Indiana University's persistent digital repository, IUScholarWorks. Such stored VM resources will be accessible via a Digital Object Identifier (DOI).
- expand access to the XSEDE ecosystem by making virtual desktop services accessible from institutions with limited resources, such as small schools, schools in Experimental Program to Stimulate Competitive Research (EPSCoR) states, and Minority Serving Institutions.

Within the XD ecosystem Jetstream will offer a fundamentally new approach to computational service delivery for the national science and engineering research community. Implementation of Jetstream will greatly increase the number of NSF-supported researchers who make use of resources within the XD ecosystem. Jetstream will address the needs of researchers who have large quantities of digital data, many of whom will benefit from the ability to do their computing within a uniquely configured virtual computing environment. Many of these researchers work in the "long tail of science" [11] and need significantly more interactive computational resources than are at their disposal on their home campus. NSF-funded resources that compose the XD ecosystem are excellent at providing scientists with thousands of processor cores in a few days or next week. If what researchers need is interactive access to a handful of processors now, whenever now is, they will not find that need well met within the XD ecosystem as of July 2015. Such interactive computing services will be available within the XD ecosystem when Jetstream becomes generally available for production science and engineering research use in 2016.

In the atmosphere, the jet stream is a fast-moving current of air at the boundary of two large air masses in the upper atmosphere. In an analogous way, the Jetstream system will serve as a point of contact between existing XD resources and users, and users new to the XD ecosystem, providing fast access to cutting-edge analysis tools.

3. HARDWARE CONFIGURATION

We will operate Jetstream in a way that is modeled after commercial cloud resources in terms of uptime and accessibility. Jetstream will comprise two geographically distributed clusters delivering production services, one at IU and the other at TACC, with a small test system at the University of Arizona. Each production instance will be connected to Internet2 via a 100 Gbps link. Jetstream will operate in at least two "zones" at TACC and IU so the system can be available for users with very high reliability, even though at times only part of the total capacity of Jetstream is available for use.

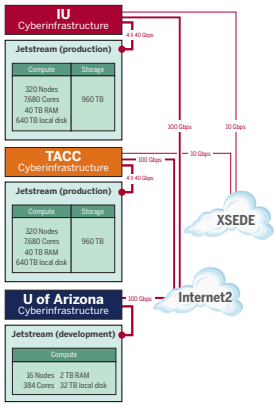

Figure 1: Schematic diagram of Jetstream physical hardware and network connectivity

Jetstream will be based on the recently released Intel 2680 v3 processor (Haswell). It is a 2.5 GHz-based processor, but operates at 2.1 GHz when executing floating-point operations that utilize the AVX (Advanced Vector Extensions) instruction set (such as Linpack). The calculations in Table 1 use the 2.1 GHz AVX base frequency. For applications that can take advantage of turbo frequencies, performance will be higher than implied by Table 1.

Table 1: Hardware specifications for Jetstream components

Jetstream component	# CPUs	# Cores	PFLOPs	Total RAM (GB)	Secondary storage (TB)	Node local storage (TB)	Connection to Internet2 (Gbps)
Production components of Jetstream							
IU	640	7,680	0.258	40,960	960	640	100
TACC	640	7,680	0.258	40,960	960	640	100
Jetstream test and build system							
Arizona	32	384	0.013	2,048	192	32	100
Total	***1,312***	***15,744***	***0.529***	***83,968***	***2,112***	***1,312***	***300***

Table 2 provides a more system-centric view of the production clusters that will be installed at IU and TACC as part of Jetstream.

Table 2. High-level hardware specifications for included as part of Jetstream

Characteristic	Production portion specifications (half in Bloomington, IN, half in Austin, TX)
Node architecture	1,280 Intel E5-2680 v3, 12-cores, 2.5 GHz (compute/head nodes)
RAM in nodes	128 GB of DDR4-2133 ECC RAM per node
Disk in nodes	2 X 1 TB SATA (7.2K RPM, 6 Gb/s 64 MB cache)
Node to chassis switch (16 nodes/switch)	10 Gbps (4:1 Fat-Tree)
Chassis to cluster core switch	40 Gbps
Networking cluster to storage and cluster to data center core routers	4 x 40 Gbps
Disk/database storage	Up to 0.96 PB total (0.48 PB per location disk storage)
Tape storage	Up to 2 PB per location

In its initial implementation Jetstream will be configured with homogeneous nodes. We will offer a variety of VM sizes, as shown in **Table 3** below.

Table 3. VM instance configurations, showing vCPUs (virtual CPUs), RAM, and storage characteristics

Instance Type	vCPUs	RAM (128 GB total)	Storage (2,048 GB total)	Instances/ Node
Tiny	1	2	20	46
Small	2	4	40	23
Medium	6	16	130	7
Large	10	30	230	4
X-Large	22	60	460	2
XXL	44	120	920	1

4. SOFTWARE ENVIRONMENT

The Jetstream software environment will be a first of a kind within the XD ecosystem. The most distinctive aspect of the Jetstream software environment will be the user interface. Jetstream will also be the first system within the XD ecosystem based on use of the OpenStack cloud software suite.

4.1 Jetstream user interface

The Jetstream user interface is shown in Figure 2. Jetstream will utilize the Atmosphere interface developed by the University of Arizona [12]. The NSF-funded iPlant Collaborative [13,14] has operated Atmosphere in production during the last 3.5 years. iPlant services and the Atmosphere interface have been widely adopted by a diverse community of life scientists with varying levels of expertise, from novice to experienced computational biologist. Through this interface a user can select from a library of pre-created VMs that perform specific analysis functions, open up a basic VM to host Linux-based applications, or create a customized VM.

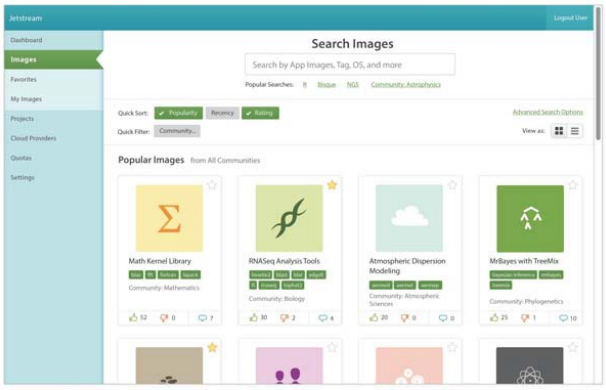

Figure 2. Planned Jetstream user interface, which is an implementation of the Atmosphere interface.

The Atmosphere interface addresses key usability issues for cloud infrastructure by: 1) integrating identity management (IDM); 2) integrating network configuration and security policies; 3) providing instance lifecycle management tools and utilities, including capabilities for metering performance, sharing, and collaboration; and 4) simplifying tools for data lifecycle management. These integrated capabilities, managed through a web portal or web services API (application programming interface), provide users with unified access for managing computation and analysis in a highly configurable cloud environment. At the same time, the interface shields users from the complexities that can be involved in using cloud platforms.

Through this interface users can choose from a variety of virtual machine images, including a public library of pre-configured VMs for specific analysis tasks and VMs created for a specific use and private to that user or shared by a small group of users. The key operations concept for Jetstream is this: Access the Jetstream Atmosphere interface, authenticate, launch a VM, and work (interactively).

Baseline VM images for Jetstream will make use of CentOS and/or Ubuntu Linux. In addition to security and resource management software, these images will feature SSH (secure shell) and VNC (virtual network console) support, plus high-speed data transfer tools such as iRODS icommands and Globus Transfer. Prebuilt VMs will offer a number of features, such as:

- bioinformatics tools, including VMs available in iPlant and Galaxy [15]
- the XSEDE-Compatible Basic Cluster build (XCBC): A VM built with the open software tools on a typical cluster in the XD ecosystem [16].
- use of proprietary software: ArcGIS [17] (usable from the IU portion of Jetstream only) and MATLAB® [18] (usable within Jetstream based on licenses possessed by end users).
- polar science: Tools for accessing data from the National Snow and Ice Data Center [19].
- network science: Open-source tools for network analysis [20].
- science gateways: A basic VM that can be used to create a science gateway Jetstream can deliver as a persistent service.
- Linux virtual desktops: For researchers and students at small schools with local personal computers and/or inadequate network connections. Limits on bandwidth will place some constraints on what can be done via virtual desktops [21].

4.2 Software stack – metal to Atmosphere

The overall software stack for Jetstream is shown in Figure 3.

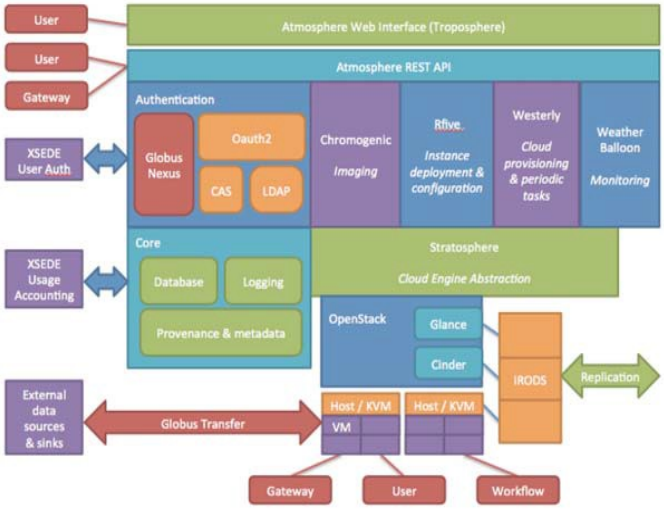

Figure 3. Schematic diagram of overall Jetstream software environment.

As part of the Jetstream implementation we will integrate Atmosphere with Globus Auth and Globus Transfer [22]. These are best-of-class technologies used extensively within XSEDE, and other diverse infrastructure sites and projects. Authentication to Jetstream will be done via Globus Auth. Bulk movement of data between Jetstream and other CI resources will be supported by Globus Transfer.

OpenStack will be the primary cloud environment powering Jetstream. OpenStack began in 2010 as a collaboration between NASA (National Aeronautics and Space Administration) and Rackspace to develop an open-source cloud, and has grown to over 1,200 contributors [23]. Like Atmosphere, it is a service-oriented architecture, allowing independent scaling of services. It orchestrates virtualized assets within the cloud, including compute, network, and storage. Key services include Glance, which holds the repository of VM images and snapshots, and Cinder, which implements block storage resources. Jetstream VMs will be provisioned using the open-source Kernel-based Virtual Machine (KVM) [24].

Atmosphere leverages the iRODS data management system [25] to store and manage entities within OpenStack such as Glance snapshots, Cinder block storage, and KVM running state images. Jetstream will use iRODS replication to keep backup copies of these files in secondary storage at the other Jetstream site. During planned or unscheduled outages, failover capabilities will leverage the backups and provide transparent high availability for services and computations.

Jetstream will use iRODS for data management within the distributed Jetstream system. Copies of all frequently used snapshots and block stores (along with paused running state images) will be kept at TACC and IU. Based on a common iRODS policy, data will be migrated from primary stores to off-system disk storage and long-term tape systems at both sites. User requests for VM images and paused instances that have been migrated to tape due to age or inactivity will be migrated back to the system when the user of those images or instances initiates a restart.

5. NEEDS DRIVING DESIGN

Jetstream was designed using a formal architectural process based on definition of use cases [26] and TOGAF (The Open Group Architecture Framework) architectural processes [27]. The design was based on a needs analysis of communities of researchers and research educators and NSF's goal of supporting the "entire spectrum of NSF-funded communities ... and to promote a more comprehensive and balanced portfolio that in turn supports the entire scientific, engineering and educational community" [28]. Our plans to promote Jetstream adoption are informed by current social-science-based understandings of technology adoption [29] that suggest 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 end users will find it accessible). Our strategy in defining Jetstream uses was to identify communities that do not currently make extensive use of XD resources, document and understand their needs and work patterns, and design a computing resource to meet those needs. One particularly important aspect of our strategy is to collaborate directly with entities that serve as aggregators of community interest and activity, such as major software projects, data resources, and groups with common needs, to create conditions that facilitate adoption of Jetstream and engage community leaders as partners to leverage their social influence.

The use cases on which we focused in designing Jetstream are described below, grouped by domain-based use cases then use cases based on modality of use.

5.1 Domain-based community use cases

We began to design Jetstream by interviewing domain scientists who need cloud-based interactive computing services for their research.

Biology (especially genomics). The iPlant Collaborative [13,14] and the Galaxy Project [15] serve and represent large biology user communities who want to do research analysis interactively, and who use cloud computing to address these needs. Demand routinely exceeds capacity of the systems on which iPlant and Galaxy are delivered. iPlant is not part of the XD ecosystem, and Galaxy is delivered from TACC systems that are not part of the XD program. *Planned solution: Provide iPlant VM and Galaxy services using Jetstream.*

Earth science/polar science – users from National Snow and Ice Data Center (NSIDC). NSIDC [19] is an aggregator of

community interests by virtue of curating and managing widely used data, but with no community collection of analysis routines. A polar researcher might know where to get data but lack access to best-practice analysis routines and sufficient computing to perform analyses. At least 2,500 researchers regularly use NSIDC-managed data products, but no common computing infrastructure is available. *Planned solution: Jetstream staff help NSIDC staff create and publish VMs that can request NSIDC data and run common data reduction and analysis routines.*

Engineering. Engineering researchers often use proprietary software, including MATLAB®. Licenses are very expensive. *Planned solution: Enable use of MATLAB® on Jetstream by enabling users to use local MATLAB licenses through their local FlexLM license manager and run MATLAB on a VM on Jetstream.*

Field station research. Biological field stations can combine data across long-term studies to facilitate new science, including studies of global climate change [30]. It's hard to maintain records on site as researchers come and go. Dr. Bryan Heidorn is creating web-accessible tools to enable researchers to upload field data others can use. The goal is a set of VM images and associated data sets supporting US field station research, including marine research at the University of Hawaii. *Planned solution: Support development of VM-based data collection and analysis tools to support field station data sharing and collaboration. Deploy VMs to support oceanographic research by the University of Hawaii field stations.*

Geographical Information Systems (GIS). Population growth, climate change, and competing land uses have created grand challenges ranging from sustainability to health and wellness. Addressing these requires advanced GIS tools. ArcGIS [17] is a popular proprietary product not currently available to the US research community in a cloud environment. *Planned solution: Provide access to ArcGIS in a VM using IU's existing site license, and deliver the CyberGIS toolkit from within a VM on Jetstream.*

Network science. Network science is the study of "network representations of physical, biological, and social phenomena leading to predictive models of these phenomena" [31]. IU School of Informatics and Computing Professor Katy Börner leads the CyberInfrastructure Shell (CIShell) effort, a software framework that five research groups use to provide analysis tools (macroscopes) for network studies [20]. Around half of CIShell tools are available in a VM-based environment, but none are installed on resources within the XD ecosystem. *Planned solution: Work with CIShell tool builders to create Jetstream VMs and deliver network analysis tools interactively.*

Social sciences. The increase in availability of "born digital" data affects the social sciences in particular, as much data was generated with embedded geographical location information. The Odum Institute [32] at the University of North Carolina is developing social science analysis and research tools to be packaged in VMs designed for the type of environment proposed for Jetstream. These VMs will allow selection of data from their Odum Institute Dataverse Network in a way that retains provenance and version information. They are especially interested in using open-source statistical software such as R [33] and GIS software such as ArcGIS. *Planned solution: Create an environment that enables the Odum Institute and other social scientists to publish analysis routines that preserve provenance.*

Text analysis and computing for humanities researchers. The HathiTrust Research Center (HTRC) [34], led by a group of researchers including IU School of Informatics and Computing Professor Beth Plale, is developing tools for large-scale analysis of textual material, where the amount of material published is beyond what any one person can assess. The National Center for Supercomputing Applications and the University of Michigan are partners. HTRC tools are suited to a cloud-oriented front end and a well secured computational backend running on large-scale clusters or supercomputers. However, HTRC lacks adequate infrastructure. *Planned solution: Deploy science gateways supporting access to and use of text analysis tools created and deployed by HTRC.*

Generalization of use cases. Researchers want straightforward access to interactive tools to analyze data, delivered in a manner congruent with their normal operations and often driven by availability of new data. A tool producer develops new analysis routines and methods to address research needs to make this functionality available to experimentalists without their having to contend with operating system complexities and software dependencies. *Planned solution: Develop an accessible platform where application creators can easily publish and share within a VM image, and end users can easily invoke runnable instances of these applications via virtualization.*

5.2 Use cases based on mode of use

In planning for Jetstream, we identified several use cases based on the mode of usage rather than the particular research interests of the user.

Campus bridging. Detailed campus bridging use cases have been published by XSEDE [35]. One has special relevance to Jetstream: the ability to initiate an interactive computing session, detach with it running, and re-attach and continue working. This would be especially useful at schools with limited CI budgets. *Planned solution: Provide a Jetstream VM image featuring a user-friendly virtual Linux desktop running on Jetstream with screen images delivered to tablet devices on cellular connections or to older PCs on slow networks. The virtual desktop will guide users to information, training materials, and XSEDE and XD program resources.* More information is available at [21].

Enable use of proprietary software. The critical path for many research analyses includes proprietary, licensed applications that will make modest use of parallelism. *Planned solution: Enable Jetstream users to run proprietary software using licenses metered through their own licenses manager.*

Facilitate reproducible data analyses. The scientific community is striving to enable reproducibility of data analyses and published research. Services like RunMyCode [36] allow distribution of software, data, and scripts, but do not provide environments in which to run the code. Commercial cloud services can be used, but at indeterminate cost and high complexity. *Planned solutions: Enable researchers to easily publish a VM image containing their analysis tools and, if desired, the input data, scripts, and output data, through a service like RunMyCode or persistent digital archive like IUScholarWorks [37].*

Enhance ease of science gateway deployment. Science gateways provide web-accessible implementations of particular analyses and scientific workflows. While straightforward to use, they can be labor intensive and often require extensive server-side programming to implement. *Planned solution: Provide a gateway builder's toolkit, including VMs with commonly used workflow engines installed and ready to configure, XSEDE tools, and a platform for persistently hosting web services.*

Visualization and analysis. Many researchers would like to interactively use visualization and analysis tools, possibly also

using large-scale XSEDE resources for visualization, data management, and simulation. Common interactive software tools include open-source or proprietary analysis tools. *Planned solution: Create a set of VMs that deliver pre-configured visualization and analysis tools to move data among XD resources, and use workflow engines to orchestrate distributed workflows between Jetstream and other XD resources.*

5.3 Jetstream capabilities and community needs

David Lifka of Cornell University (a Jetstream project Senior Investigator) and colleagues surveyed a large number of projects using cloud computing resources in order to assess the types of cloud services that XSEDE should be positioned to support [38]. Jetstream will support 10 of the 12 use cases identified in this report. We do not plan to address two types of needs identified in the report: clouds for computer science research and "bursting" of compute jobs to cloud facilities. Support for computer science research on cloud environments is the primary function of the new Chameleon [39] and CloudLab systems [40], successors to the IU-led FutureGrid [41,42]. We expect that the overall usage levels of Jetstream will be so high that the system will not often have sufficient idle capacity to support bursting of compute jobs to Jetstream.

There are other potential needs we may address in the future that we will not address in Jetstream's initial deployment. Most notably we hope to add Docker support after the system has been accepted by the NSF and is available for general use in 2016.

6. RELATIONSHIPS BETWEEN WRANGLER, iPLANT, AND OTHER PROJECTS

Wrangler is a new high-speed storage and analytics resource designed for large-scale data transfer, analytics, and sharing. Wrangler will provide flexible support for a wide range of software stacks and workflows [5]. TACC leads the Wrangler project and IU is a partner in that effort. Wrangler's 20 PB of spinning disk will be split between TACC and IU, with 10 PB at each site. With cloud environments split between IU and TACC, Jetstream's implementation suggests natural synergy opportunities between Wrangler and Jetstream. Jetstream will have sufficient disk storage capacity to support its users. The opportunity to use community data sets maintained on Wrangler will open new possibilities for the US research community. The Wrangler analytics cluster located at TACC will also complement Jetstream, which is configured primarily for jobs that require a modest number of processors and VMs simultaneously in use. The Wrangler analytics cluster will be better suited than Jetstream for workloads focused on very large MapReduce jobs and NoSQL database usage. Jetstream and Wrangler will together provide a set of cloud services that will meet a wide variety of US open research needs.

There is by design overlap between the Jetstream construction partnership and other major projects that involve TACC, PTI, the University of Arizona, and Johns Hopkins University. Jetstream is distinctive within the XD ecosystem in terms of the number of partners involved in delivering a computational and data analysis system. Several partners will be involved in delivering software via Jetstream and providing outreach and dissemination of information about Jetstream within science and engineering communities. Once the system has been accepted by the NSF and put in to production in 2016, multiple organizations will be involved in Jetstream maintenance and operation, including:

- Cornell University: creation of virtual workshops for training and outreach.
- Johns Hopkins University and Penn State University: lead institutions in the Galaxy community, responsible for leading assessment of priorities in implementing and supporting Galaxy on Jetstream.
- National Snow and Ice Data Center, lead institution for ice-sheet data analysis and partner in deploying tools to use data housed at NSIDC.
- Odum Institute: lead institution for social sciences support and source of VMs deployed for analysis on Jetstream.
- University of Arkansas Pine Bluff: cybersecurity education and outreach to Historically Black Colleges and Universities (HBCUs) and EPSCoR states.
- University of Texas San Antonio Cloud and Big Data Lab: OpenStack support and Hispanic Serving Institution outreach.
- University of Hawaii: oceanography applications and EPSCOR state outreach (Dr. Gwen Jacobs of the University of Hawaii will chair the Jetstream Stakeholder Advisory Committee).

7. CHALLENGES IN IMPLEMENTATION

Any first-of-a-kind system like Jetstream presents multiple challenges: for the Jetstream team and for the NSF in defining acceptance criteria, for XSEDE and the XSEDE Resource Allocation Committee (XRAC) in allocating the system, and for XSEDE and the Jetstream team in supporting the system.

The NSF will be breaking new ground in its acceptance testing for Jetstream. The challenge is to define a set of tests that demonstrate the functioning of the Jetstream subunits in Indiana and Texas as an integrated cloud system. This is also a challenge for our vendor partner, Dell, which has gracefully accepted that the Jetstream team is the system integrator. Dell is dependent on software implementation beyond its direct control for system acceptance by the NSF. Payment for acquisition of the system will not be available for Dell until the NSF declares acceptance. The Jetstream team and NSF will set the precedents for how the NSF accepts production-oriented cloud systems (as opposed to experimental cloud systems such as FutureGrid, Chameleon, and CloudLab).

As the first NSF production cloud computing resource, Jetstream brings new challenges to the process for allocating time on the system. Our goal is that getting an account on Jetstream will happen with much the same speed as getting an account on a commercial cloud system. Our proposed allocations process is similar to the iPlant process, implemented in ways that are consistent with NSF guidance and existing XSEDE allocation practices. Our planned approach will recognize that work styles of computational researchers using cloud systems are very different from those of researchers using large-scale HPC systems. We have proposed three variations on the XSEDE "startup allocations" concept, as follows:

- An *automatic* startup allocation: Highly constrained to let researchers see if Jetstream will fulfill their needs. Requires only an XSEDE User Portal (XUP) account. An automatic startup is created automatically on request, with a single VM with low resource consumption and tight security constraints.

- A *traditional* XSEDE startup allocation: For researchers who need more than the single VM. A traditional one-year allocation with additional resources requested/granted.
- A *hybrid* allocation: For researchers who want a small extra amount of resource—the target Jetstream demographic of the thousands of researchers not currently served by NSF HPC. We have proposed that researchers request resources directly from the Jetstream project team, which is authorized to grant allocations larger than the standard startup. This seems the best scalable way to reach these thousands of researchers.

These allocations types are consistent with the existing XSEDE approach for startup allocations. Startup allocations are acted on by staff members, without direct XRAC involvement unless the amount of resource requested merits such review. For researchers who require more than startup allocations, Jetstream will utilize the normal XRAC process and existing XRAC review and allocation process.

Jetstream support will be a challenge for Jetstream and XSEDE. The existing formula for maintenance and operations of an NSF-funded XD ecosystem resource is based on the assumption that XSEDE provides significant assistance in account management, documentation, advanced user support, and education and outreach. There is not enough money in the maintenance and operations budget for Jetstream to do these functions, and XSEDE so far has no experience supporting a cloud resource used by the general scientific community. We plan to leverage existing funding to our partner organizations, but we expect that the scale of the Jetstream user community will require adjustments to XSEDE and NSF plans for delivering user support.

Capacity and managing appropriate use may also be challenging. Our success in proposing Jetstream to the NSF is partly based on incorporating into the project team pre-existing communities that needed the resource. Once Jetstream becomes available for production use in 2016 we expect to have thousands of Jetstream users relatively quickly who are new to the XD ecosystem. The more successful we are in attracting users, the more quickly we will hit capacity limitations. The system will be horizontally scalable. Software and application integration are the hard parts; expanding capacity will be straightforward given additional federal funding.

We also expect some degree of challenge in managing appropriate use of Jetstream. Campus leaders have sometimes said they appreciate NSF-funded cyberinfrastructure resources because these resources allow campuses to move the cost of delivering CI resources from the campus to federally funded facilities. That's not the purpose of the XD ecosystem as defined in the several NSF solicitations that have resulted in funding for the current set of resources. We will need to take care that use of resources via Jetstream is not an alternative to institutions investing in local CI resources.

8. DISCUSSION AND CONCLUSION

As a first-of-a-kind cloud system, Jetstream will support new types of interactive CI use for the NSF-funded XD ecosystem. Users will largely complement those who use the existing XD ecosystem. Jetstream will greatly expand XD ecosystem value to the national research community, notably in disciplines and sub-disciplines in NSF-supported areas of research and engineering.

In the last two years the NSF has invested in systems that support innovative modes of computing such as Wrangler and Comet. Investment in the Pittsburgh Supercomputing Center's Bridges system provides new resources for the XD user community while engaging in new forms of campus bridging. Investment in Jetstream will add a general-purpose science and engineering cloud resource to the XD ecosystem.

While many challenges lie ahead for XSEDE and the Jetstream team in deploying this new service, Jetstream should accelerate US research achievements, support research in the long tail of science, allow for new reproducibility in computationally based research, and open up new avenues for research and research education in institutions with limited CI budgets, such as Minority Serving Institutions and some institutions in EPSCoR states.

The NSF's investment in Jetstream and Bridges will support new and existing users of the XD ecosystem, and aid US research productivity and global competitiveness.

9. ACKNOWLEDGMENTS

Implementation of Jetstream is supported by NSF award 1445604. The Indiana University Pervasive Technology Institute was established with support of funding from the Lilly Endowment, Inc., and provided support for this research. Any opinions expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation, the Lilly Endowment, or XSEDE leadership as a whole. We thank XSEDE leadership and staff and NSF staff for thoughtful discussion and assistance in making this report possible, and three anonymous reviewers whose comments helped us improve this final report.

10. REFERENCES

[1] Towns, J., T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G.D. Peterson, R. Roskies, J.R. Scott and N. Wilkins-Diehr. 2014. XSEDE: Accelerating Scientific Discovery. In Computing in Science & Engineering, 16(5): 62-74. doi:10.1109/MCSE.2014.80

[2] XSEDE. Home page. https://www.xsede.org/

[3] XSEDE. Requesting Membership in the XSEDE Federation, V2, 10 July 2014. http://hdl.handle.net/2142/49981

[4] Moore, R.L., C. Baru, D. Baxter, G. Fox, A. Majumdar, P. Papadopoulos, W. Pfeiffer, R.S. Sinkovits, S. Strande, M. Tatineni, R.P. Wagner, N. Wilkins-Diehr, M.L. Norman. Gateways to Discovery: Cyberinfrastructure for the Long Tail of Science. In Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment. 2014. ACM: Atlanta, GA, USA. p. 1-8.

[5] Texas Advanced Computing Center. Wrangler. https://www.tacc.utexas.edu/-/wrangler-data-intensive-system-opens-to-scientists

[6] Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020. 2014. Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020: Interim Report. Washington, DC. The National Academies Press. 48 pp. http://www.nap.edu/catalog/18972/future-directions-for-nsf-advanced-computing-infrastructure-to-support-us-science-and-engineering-in-2017-2020

[7] Pittsburgh Supercomputing Center. Bridging the GAP. 2014. Available from: http://www.psc.edu/index.php/bridging-the-gap

[8] Nystrom, N.A., M.J. Levine, R.Z. Roskies, and J.R. Scott. 2015. Bridges: A Uniquely Flexible HPC Resource for New Communities and Data Analytics. In: Proceedings: XSEDE'15, July 26 - 30, 2015, St. Louis, MO, USA. Doi: http://dx.doi.org/10.1145/2792745.2792775

[9] Indiana University. 2014. Jetstream. http://www.jetstream-cloud.org

[10] Indiana University Pervasive Technology Institute. Home Page. http://pti.iu.edu/

[11] Heidorn, P.B. 2008. Shedding Light on the Dark Data in the Long Tail of Science. Library Trends 57(2), p. 280-299. http://muse.jhu.edu/journals/library_trends/v057/57.2.heidorn.html

[12]Skidmore, E., S.-j. Kim, S. Kuchimanchi, S. Singaram, N. Merchant, and D. Stanzione. 2011. 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. http://dl.acm.org/citation.cfm?id=2110495

[13] iPlant Collaborative. Home Page. http://iplantcollaborative.org/

[14] Goff, S.A., M. Vaughn, S. McKay, E. Lyons, A.E. Stapleton, D. Gessler, N. Matasci, L. Wang, M. Hanlon, A. Lenards, A. Muir, N. Merchant, S. Lowry, S. Mock, M. Helmke, A. Kubach, M. Narro, N. Hopkins, D. Micklos, U. Hilgert, M. Gonzales, C. Jordan, E. Skidmore, R. Dooley, J. Cazes, R. McLay, Z. Lu, S. Pasternak, L. Koesterke, W.H. Piel, R. Grene, C. Noutsos, K. Gendler, X. Feng, C. Tang, M. Lent, S.-J. Kim, K. Kvilekval, B. S. Manjunath, V. Tannen, A. Stamatakis, M. Sanderson, S.M. Welch, K.A. Cranston, P. Soltis, D. Soltis, B. O'Meara, C. Ane, T. Brutnell, D.J. Kleibenstein, J.W. White, J. Leebens-Mack, M.J. Donoghue, E.P. Spalding, T.J. Vision, C.R. Myers, D. Lowenthal, B.J. Enquist, B. Boyle, A. Akoglu, G. Andrews, S. Ram, D. Ware, L. Stein, and D. Stanzione. 2011. The iPlant Collaborative: Cyberinfrastructure for Plant Biology. In Frontiers in Plant Science. http://journal.frontiersin.org/article/10.3389/fpls.2011.00034/abstract

[15] Goecks, J, A. Nekrutenko, J. Taylor, and The Galaxy Team. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11(8):R86. http://genomebiology.com/2010/11/8/R86

[16] Fischer, J., R. Knepper, M. Standish, C.A. Stewart, R. Alvord, D. Lifka, B. Hallock, and V. Hazlewood. 2014. Methods For Creating XSEDE Compatible Clusters. In XSEDE '14: Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment. 13-18 July, Atlanta, GA. ACM. http://dl.acm.org/citation.cfm?id=2616578

[17] ArcGIS. Home Page. https://www.arcgis.com/

[18] Mathworks. Home Page. http://www.mathworks.com

[19] National Snow and Ice Data Center. Home Page. http://nsidc.org

[20] NWB Team. (2006). Network Workbench Tool. Indiana University, Northeastern University, and University of Michigan, http://nwb.slis.indiana.edu

[21] Fischer, J., S. Tuecke, I. Foster, C.A. Stewart. 2015. Jetstream: A Distributed Cloud Infrastructure for Underresourced higher education communities. In SCREAM '15: Proceedings of the 1st Workshop on The Science of Cyberinfrastructure: Research, Experience, Applications and Models. pp 53-61. ACM New York, NY. http://dl.acm.org/citation.cfm?doid=2753524.2753530

[22] Globus. Home Page. https://www.globus.org/

[23] OpenStack. Home Page. http://www.openstack.org

[24] KVM. Home Page. http://www.linux-kvm.org/page/Main_Page

[25] iRODS. Home Page. http://www.irods.org

[26] Malan, R. and D. Bredemeyer. 2001. Functional Requirements and Use Cases, in Architecture Resources for Enterprise Advantage. www.bredemeyer.com/pdf_files/functreq.pdf

[27] The Open Group. 2014. TOGAF® Version 9.1. http://www.opengroup.org/togaf/

[28] National Science Foundation. 2014. High Performance Computing System Acquisition: Continuing the Building of a More Inclusive Computing Environment for Science and Engineering Program Solicitation NSF 14-536. http://www.nsf.gov/pubs/2014/nsf14536/nsf14536.htm

[29] Venkatesh, V., Morris, M.G., Davis, F.D., Davis, G.B. 2003. User Acceptance of Information Technology: Toward a Unified View. MIS Quarterly, 27(3), 425-478. https://csdl-techreports.googlecode.com/svn/trunk/techreports/2005/05-06/doc/Venkatesh2003.pdf

[30] Wyman, R.L., E. Wallensky, M. Baine. 2009. The Activities and Importance of International Field Stations. BioScience, 59(7): p. 584-592. http://bioscience.oxfordjournals.org/content/59/7/584.full

[31] Committee on Network Science for Future Army Applications, 2005. Network Science. The National Academies Press. http://www.nap.edu/openbook.php?record_id=11516

[32] Howard W. Odum Institute for Research in Social Science. Home Page. http://www.odum.unc.edu/odum/home2.jsp

[33] The R Project for Statistical Computing. Home Page. http://www.r-project.org/

[34] HathiTrust Research Center. Home Page. http://www.hathitrust.org/htrc

[35] Stewart, C.A., R. Knepper, J.W. Ferguson, F. Bachmann, I. Foster, A. Grimshaw, V. Hazlewood, D. Lifka. 2012. What is campus bridging and what is XSEDE doing about it? In Proceedings of XSEDE12. July 16-19, Chicago, IL. ACM New York, NY. Doi: 10.1145/2335755.2335844

[36] RunMyCode. Home Page. http://www.runmycode.org/

[37] Indiana University. IUScholarWorks. http://scholarworks.iu.edu/

[38] Lifka, D., I. Foster, S. Mehringer, M. Parashar, P. Redfern, C.A. Stewart, and S. Tuecke, 2013. XSEDE Cloud Survey Report. 2013. http://www.cac.cornell.edu/technologies/XSEDECloudSurveyReport.pdf

[39] Chameleon. Home Page. https://www.chameleoncloud.org/

[40] Cloudlab. Home Page. http://www.cloudlab.us/

[41] FutureGrid. Home Page. http://futuregrid.org/

[42] Fox, G.C., G. v. Laszewski, J. Diaz, K. Keahey, J. Fortes, R. Figueiredo, S. Smallen, W. Smith, and A. Grimshaw. 2013. FutureGrid - a reconfigurable testbed for Cloud, HPC, and Grid Computing. In J. Vetter (ed). Contemporary High Performance Computing: From Petascale toward Exascale. Chapman & Hall. https://portal.futuresystems.org/references/futuregrid-reconfigurable-testbed-cloud-hpc-and-grid-computing