Jetstream A Cloud System Enabling Learning in Higher Education Communities

SIGUCCS17 – October 3, 2017 – Seattle, WA

Jeremy Fischer – jeremy@iu.edu
Senior Technical Advisor,
UITs Research Technologies

NSF Funding Areas in HPC

Traditionally concentrated on enabling petascale capability
- Blue Waters – 13.3 petaflops, 2012
- Stampede – 9.6 petaflops, 2013
- Comet – ~2.0 petaflops, 2014

Has funded research into building clouds and computer science
- CloudLab
- Chameleon

Now funding clouds to do research
- Bridges (Hybrid system)
- Jetstream
Lots of stats below –

**tl;dr summary: no one has enough computing resources. Ever. But they need easy access and use.**

Around 350,000 researchers, educators, & learners received NSF support in 2015

- Less than 2% completed a computation, data analysis, or visualization task on XD/XSEDE program resources
- Less than 4% had an XSEDE Portal account
- 70% of researchers surveyed* claimed to be resource constrained

Why are the people not using XD/XSEDE systems not using them?

- Perceived ease of access and use
- HPC resources – the traditional view of what XSEDE offers - are often not well-matched to their needs
- They just don’t need *that much* capability

* XSEDE Cloud Survey Report - [http://hdl.handle.net/2142/45766](http://hdl.handle.net/2142/45766)
What is Jetstream and why does it exist?

- NSF’s first production cloud facility
- Part of the NSF eXtreme Digital (XD) program
- Provides on-demand *interactive* computing and analysis
- Enables *configurable* environments and *programmable cyberinfrastructure*
- User-friendly, widely accessible cloud environment
- User-selectable library of preconfigured virtual machines

funded by the National Science Foundation
Award #ACI-1445604
What is Jetstream, continued…

- Focus on ease-of-use, broad accessibility
- Command line access for those who want it and GUI access for those who don’t
- Will support persistent gateways (SEAGrid, Galaxy, GenApp NAMDRunner, CIPRES and others)
- Reproducibility: Share VMs and then store, publish via IU Scholarworks (DOI)
Who uses Jetstream?

• The researcher needing a handful of cores (1 to 44/vCPU)
• Software creators and researchers needing to create their own customized virtual machines and workflows
• Science gateway creators using Jetstream as either the frontend or processor for scientific jobs
• STEM Educators teaching on a variety of subjects
Jetstream System Overview

- IU Cyberinfrastructure
  - Jetstream (production)
    - Compute: 320 Nodes, 7,680 Cores, 40 TB RAM, 640 TB local disk
    - Storage: 960 TB

- TACC Cyberinfrastructure
  - Jetstream (production)
    - Compute: 320 Nodes, 7,680 Cores, 40 TB RAM, 640 TB local disk
    - Storage: 960 TB

- U of Arizona Cyberinfrastructure
  - Jetstream (development)
    - Compute: 16 Nodes, 2 TB RAM, 384 Cores, 32 TB local disk

funded by the National Science Foundation
Award #ACI-1445604
Platform Overview

Web App

Globus Auth

Atmosphere API

Atmo Services

XSEDE Accounting

OpenStack

Ceph

OpenStack

Ceph

Indiana University

TACC
Hardware and Instance "Flavors"

VM Host Configuration

- Dual Intel E-2680v3 “Haswell”
- 24 physical cores/node @ 2.5 GHz (Hyperthreading on)
- 128 GB RAM
- Dual 1 TB local disks
- 10GB dual uplink NIC
- Running KVM Hypervisor

- Short-term *ephemeral* storage comes as part of launched instance
- Long-term storage is XSEDE-allocated
- Implemented as OpenStack Volumes
- Each user can get 10 volumes up to 500GB total storage*

<table>
<thead>
<tr>
<th>Flavor</th>
<th>vCPUs</th>
<th>RAM</th>
<th>Storage</th>
<th>Per Node</th>
</tr>
</thead>
<tbody>
<tr>
<td>m1.tiny</td>
<td>1</td>
<td>2</td>
<td>8</td>
<td>46</td>
</tr>
<tr>
<td>m1.small</td>
<td>2</td>
<td>4</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>m1.medium</td>
<td>6</td>
<td>16</td>
<td>60</td>
<td>7</td>
</tr>
<tr>
<td>m1.large</td>
<td>10</td>
<td>30</td>
<td>60</td>
<td>4</td>
</tr>
<tr>
<td>m1.xlarge</td>
<td>24</td>
<td>60</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>m1.xxlarge</td>
<td>44</td>
<td>120</td>
<td>60</td>
<td>1</td>
</tr>
<tr>
<td>s1.large**</td>
<td>10</td>
<td>30</td>
<td>120</td>
<td>4</td>
</tr>
<tr>
<td>s1.xlarge**</td>
<td>24</td>
<td>60</td>
<td>240</td>
<td>2</td>
</tr>
<tr>
<td>s1.xxlarge**</td>
<td>44</td>
<td>120</td>
<td>480</td>
<td>1</td>
</tr>
</tbody>
</table>

** s1.* based instances are not eligible to be saved into a customized image
Using Jetstream VMs

Manipulating Jetstream VMs:
• Jetstream Atmosphere web interface
• Direct API access via OpenStack command line or Horizon access
  - API access enables Science Gateways and other always on services or on demand use cases; e.g. elastic compute techniques

Primary methods of logging into Jetstream VMs to work
• Interactive user access via web interface with VNC/SSH
• Direct VNC/SSH to individual instances
The Jetstream Atmosphere web interface
Atmosphere...launching!
The instance information screen
The web shell

Jetstream
http://jetstream-cloud.org/
The web desktop
HPC vs Cloud

Adapting to a different environment:

• No reservations, no queueing
• More interactive use and less/no batch queuing
• What? No parallel filesystem?!?
• Being your own admin – hey, we have root!
• You really can have almost any (linux) software you want**
• Constantly getting new features (https://www.openstack.org/software/project-navigator/)

** Here there be dragons…
Thinking about VMs…

Cows, not pets: pets take great amount of care, feeding, and you name them; cows you intend to have high turnover and you give them numbers.

-- Mike Lowe (Jetstream architect)
Jetstream for engineering researchers (and others)

Matlab and 52 standard toolkits are installed on Jetstream.

You do NOT need to have a local license to use MATLAB on Jetstream.

If you are a researcher, and MATLAB and one of the standard first 52 toolkits that come with MATLAB help you… you’re ready to go!

If you are an engineering researcher, and you need other tools… Let us know – we are happy to consider other requests.

http://jetstream-cloud.org/

funded by the National Science Foundation
Award #ACI-1445604
Another Use Case: Galaxy riding Jetstream

Galaxy is a platform for biomedical research, focused on accessibility, transparency and reproducibility

- The main project instance (usegalaxy.org) has more than 100,000 registered users executing 300,000+ jobs each month
- Many users need more capacity than the public quota, or other customizations (e.g., new tools)

Use Jetstream as a *bursting* platform

- From Galaxy Main, offload jobs onto a remote Slurm cluster running on Jetstream instances
- Run Galaxy Interactive Environments (i.e., Dockerized IPython/RStudio containers) in an isolated environment on a Swarm cluster running on Jetstream

Use Jetstream as a *self-service* platform

- Pre-built Galaxy image configured with hundreds of tools and access to TBs of genomic reference data, available via the self-launch model within minutes
- Allows users to acquire (free) resources, and gives them complete control
Not just the usual suspects...

Physics, chemistry, and other “usual” HPC suspects are represented, but Jetstream also is home to projects on:

- Financial analysis / Economics
- Political science
- Humanities / Text analysis
- Network analysis
- Computer Science / Machine learning
- Satellite data analysis
Jetstream for Education

- Research is a primary function
  - Training current and future researchers is crucial
    - Consistent environments
    - Customizable environments
    - Easy to access (Location independent, Common tools)
    - Easy to use (Familiar environments)

http://jetstream-cloud.org/

funded by the National Science Foundation
Award #ACI-1445604
Surveying Educators on Jetstream

- Selected from *active* education allocations (mostly or entirely used)
- Selection of university courses (graduate and undergraduate) and workshops
- IRB-approved instrument for contacting PIs and key personnel
- 9 individuals (64.29%) agreed to participate with informed consent and have their results disclosed
- Only four cases are discussed due to individuals consenting to participate but requesting no disclosure, and/or allowing disclosure but not providing enough detail either through survey or follow-up to detail.
Courses covered in survey results

<table>
<thead>
<tr>
<th>Event Title</th>
<th>Type</th>
<th>Number of Students</th>
<th>Location</th>
<th>Audience</th>
</tr>
</thead>
<tbody>
<tr>
<td>Analysis of large, complex biobehavioral, bioinformatic, and genomic data sets</td>
<td>Workshop</td>
<td>36</td>
<td>Brandeis University</td>
<td>Graduate students</td>
</tr>
<tr>
<td>Bioinformatics: Tools for Genome Analysis</td>
<td>Course</td>
<td>9</td>
<td>Online (via Johns Hopkins University)</td>
<td>Graduate students</td>
</tr>
<tr>
<td>‘Digital Pedagogy’ / ‘Introduction to Text Analysis’</td>
<td>Course</td>
<td>30</td>
<td>University of Pittsburgh / Carnegie Mellon University</td>
<td>Graduate students</td>
</tr>
<tr>
<td>Management, Access, and Uses of Big and Complex Data</td>
<td>Course</td>
<td>200</td>
<td>Indiana University</td>
<td>Undergraduate and graduate students</td>
</tr>
<tr>
<td>Data-Driven Neuroimaging</td>
<td>Workshop</td>
<td>30</td>
<td>University of California San Francisco</td>
<td>Researchers and graduate students</td>
</tr>
<tr>
<td>2017 Metagenomics Workshop</td>
<td>Workshop</td>
<td>30</td>
<td>University of California Santa Cruz</td>
<td>Researchers, graduate students, and undergraduates</td>
</tr>
</tbody>
</table>
Key Points of Success

- 100% of Jetstream education PIs surveyed reported their use of Jetstream to be a success in their opinion

- “Jetstream went more smoothly than [commercial provider] in almost every way and seems to perfectly meet our need for training!”

- “fast and flexible access to computational resources at no cost”

- Beneficial to researchers to use in workshops because they could apply for their own allocations afterward and continue using the system that they learned on

- “giving students and instructors the ability to modify, install software on their own and maintain access on demand later on to those images is something of great value, which would not have been easily achieved on our cluster.”
Where can Jetstream improve?

Constructive criticism is necessary!

- Better monitoring of usage by students (workarounds now, but better options coming)
- Improve training materials
- Add videos to training materials
- Persistent IPs in Atmosphere GUI environment
- Improved base images for areas of domain science and instruction
Engaging educators

Common message of “This is great! More people should be using it! Get the word out!” (but no ideas on how to best do that)

- Engage and train IT support staff
- Presenting at conferences (PAG, SC, PEARC, SIGUCCS, targeting domain science conferences)
- Educause and regionals
- Site visits for tutorials
- Other suggestions certainly welcome!
More usage after the survey...

- Supported a month-long workshop in Summer 2017 with 70 users (second workshop for PI!)
- Currently supporting multiple semester-long courses
- Education allocation requests are rising consistently (not necessarily quickly, but consistently – numbers coming up!)
Conclusions - Jetstream for education

- Successful so far, though some hiccups
- Valuable for creating custom, uniform environments
- Must find ways to increase adoption and user base
- More accessible than commercial providers for the long run, so maybe more of a value to resource-constrained facilities
Requesting access to Jetstream

• You can request startup allocations anytime. (Startups are simple!)
• You can request allocations for educational use anytime.
• We are happy to help you prepare a request and create a successful proposal.
• You do not have to have prior use of Jetstream to be successful.
Allocation types and docs needed for each

• Startup allocation (apply anytime)
  - Current CV for PI and any Co-Pis
  - Brief abstract/description of work

• Education allocation (apply anytime)
  - Current CV for PI and any Co-PIs
  - Syllabus/Class/Workshop description
  - Description of use --> justification of SUs requested

• Research allocation (quarterly allocation window)
  - Current CV for PI and any Co-PIs
  - Main project description (up to 10 pages unless > 15M SUs, then 15 pages)
  - Scaling doc (up to 5 pages)
Jetstream Overall Highlights

As of September 1, 2017:

• 322 active XSEDE projects covering 59 fields of science and 2000+ active users representing 189 institutions
• Over 70 million CPU hours allocated to XSEDE projects since June 2016
• 9 active science gateways
• 40 education/teaching allocations serving over 600 undergraduate and graduate students
Jetstream Timeline...what comes next?

- Completed our first year of operations on September 1, 2017
- Soliciting Research allocation requests plus Startup and Education allocations – including Science Gateways!
- Adding services as deemed useful/mature (Heat, Magnum, Trove, Manila, etc)
- Atmosphere enhancements on a regular cycle
- Working on partnerships with groups like HubZero and others to extend the value of Jetstream
Where can I get help?


User guides: https://portal.xsede.org/user-guides

XSEDE KB: https://portal.xsede.org/knowledge-base

Email: help@xsede.org

Campus Champions: https://www.xsede.org/campus-champions

Training Videos / Virtual Workshops (TBD)
Jetstream Fun: Happy cluster / Angry Cluster

funded by the National Science Foundation
Award #ACI-1445604

http://jetstream-cloud.org/
Questions?

Project website: http://jetstream-cloud.org/
Project email: help@jetstream-cloud.org Direct email: jeremy@iu.edu

License Terms

• Jetstream is supported by NSF award 1445604 (Craig Stewart, IU, PI)
• XSEDE is supported by NSF award 1053575 (John Towns, UIUC, PI)
• This research was supported in part by the Indiana University Pervasive Technology Institute, which was established with the assistance of a major award from the Lilly Endowment, Inc. Opinions presented here are those of the author(s) and do not necessarily represent the views of the NSF, IUPTI, IU, or the Lilly Endowment, Inc.
• Items indicated with a © are under copyright and used here with permission. Such items may not be reused without permission from the holder of copyright except where license terms noted on a slide permit reuse.
• Except where otherwise noted, contents of this presentation are copyright 2015 by the Trustees of Indiana University.
• This document is released under the Creative Commons Attribution 3.0 Unported license (http://creativecommons.org/licenses/by/3.0/). This license includes the following terms: You are free to share – to copy, distribute and transmit the work and to remix – to adapt the work under the following conditions: attribution – you must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work). For any reuse or distribution, you must make clear to others the license terms of this work.