

Big Data, Big Red II, Data  
Capacitor II, Wrangler,  
Jetstream, and Globus Online



funded by the National Science Foundation  
Award #ACI-1445604

# Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online

**Craig A. Stewart, Ph.D. (Wittenberg class of '81)**

Orcid ID: 0000-0003-2423-9019

Executive Director, Pervasive Technology Institute

Associate Dean, Research Technologies

Indiana University

Please cite as: Stewart, C.A. 2014. Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online. Presented to Microsoft, Inc. visiting group. Indiana University, Bloomington, IN. 24 Feb 2015. <http://hdl.handle.net/2022/19685>



License specifics in last slide



Award #1445604



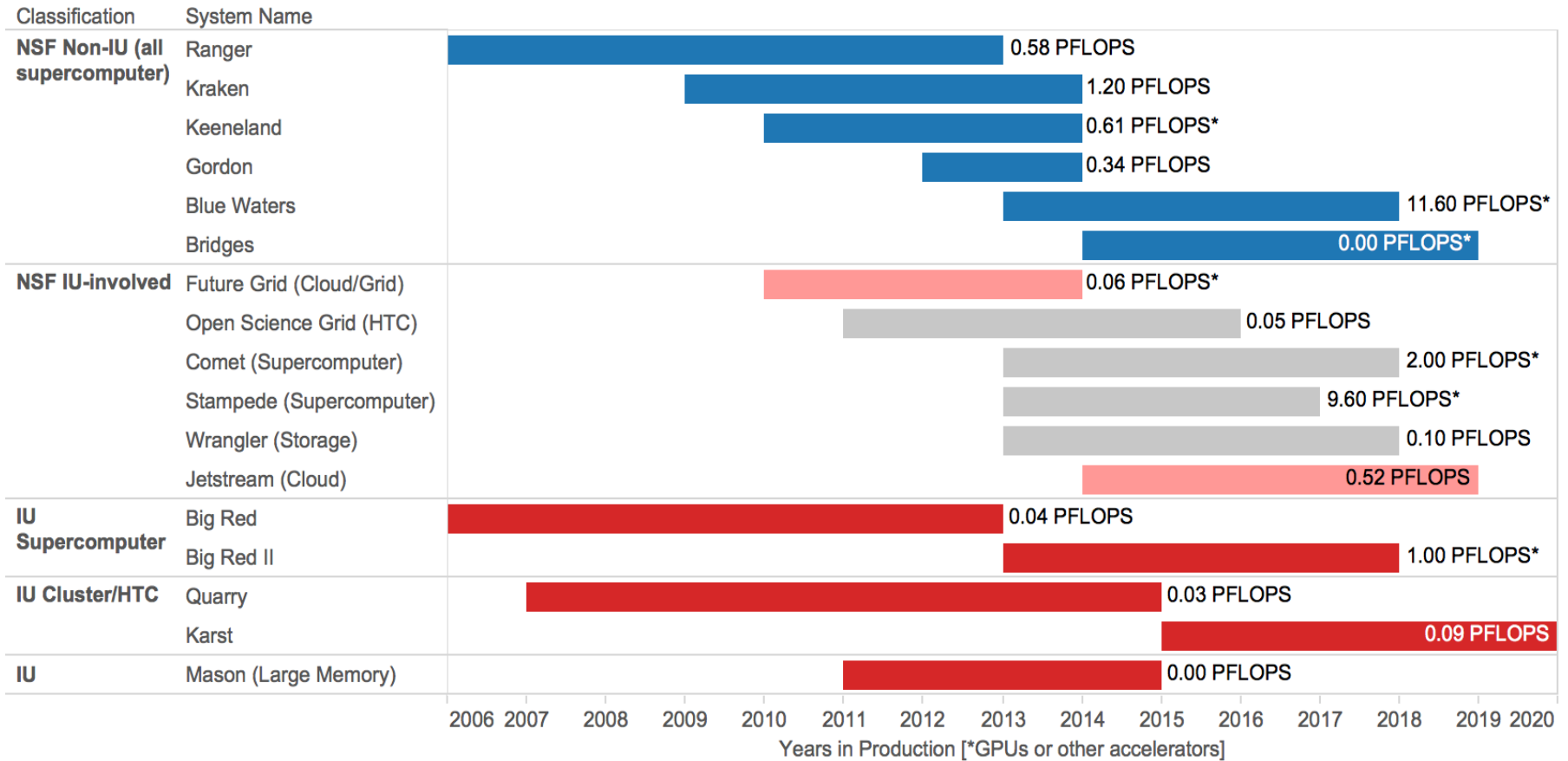
[pti.iu.edu/jetstream](http://pti.iu.edu/jetstream)

# An Intro to the UITS Research Technologies Division of UITS and Pervasive Technology Institute

- The mission of Research Technologies ... is to develop, deliver and support advanced technology solutions that improve the productivity of and enable new possibilities in research, scholarly endeavors, and creative activity at Indiana University and beyond; and to complement this with education and technology translation activities to improve the quality of life of people in Indiana, the nation, and the world.
- **RT is a mission- and value-driven organization. We are not a technology-driven organization.**
- PTI is a collaborative organization within IU involving OVPIT, UITS, SOIC, Maurer School of Law, and the College of Arts and Sciences. It puts the 'basic CS and Informatics' research at the front end of 'Discover, Develop, Deploy, Deliver, Support' for research
- PTI and RT identify needs, identify possibilities, and discover new ways to meet those needs, realize those possibilities, and create new ones. In so doing, we create, deploy, and support technology. **We are technology-driving organizations.**



# A bit about the world of big systems



## Legend

- IU System
- NSF IU Lead
- NSF IU Partner
- NSF Non-IU

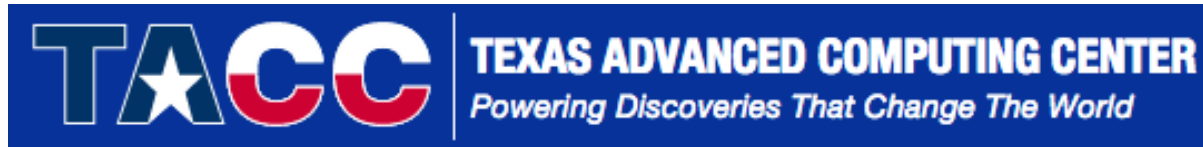


Award #1445604



[pti.iu.edu/jetstream](http://pti.iu.edu/jetstream)

# Wrangler



- Geographically replicated, high performance data storage (10PB each site)
- Large-scale flash storage tier for analytics with bandwidth of 1TB/s and 250M IOPS
- More than 3,000 embedded processor cores for data analysis
- Wide range of software stacks, including Hadoop®, R, relational and noSQL databases
- Globus services for rapid, reliable data transfer, sharing, and publication
- Partnership of Texas Advanced Computing Center, IU Pervasive Technology Institute, University of Chicago, Dell



# Jetstream

- Geographically Distributed Cloud, 0.5 PetaFLOPS
- High-speed connections to Internet2 and local connections to Wrangler disk storage at IU and TACC
- Globus-based large scale file movement



# Jetstream will . . .

- be NSF's first cloud for science and engineering research across all areas of activity supported by the NSF.
- be a user-friendly cloud environment designed to give researchers and research students access to interactive computing and data analysis resources "on demand"
- leverage Globus tools for data movement and authentication.
- provide a user-selectable library of virtual machines from which users can select to do their research.
- enable software creators and researchers to create their own customized virtual machines or their own "private computing system" within Jetstream.
- ... and store them in IUScholarWorks with a DOI for publication and support of replicability of research.
- enable discoveries across disciplines such as biology, atmospheric science, economics, network science, observational astronomy, and social sciences.



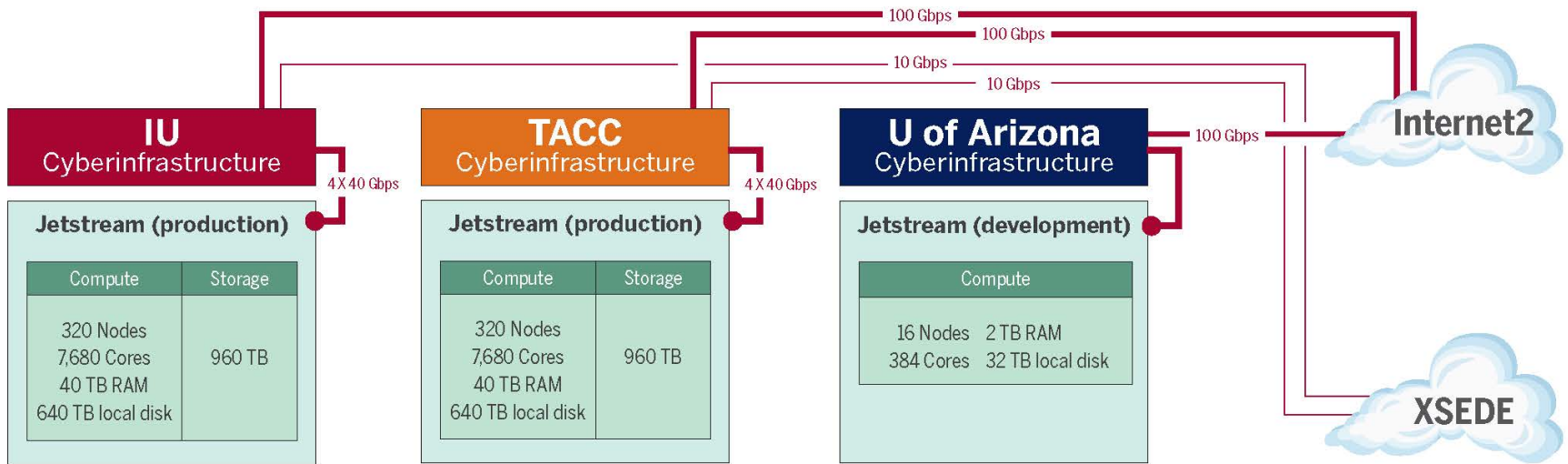
# What does the name mean? And is it really a cloud?

- Name
  - In the atmosphere the Jetstream lies at the border of two different air masses
  - The Jetstream system stands at the border of the existing NSF-funded XD program and advanced cyberinfrastructure resources and users who have not previously used such NSF funded infrastructure.
- Yep, it's really a cloud, or at least a cloud environment (one could quibble over the definition of cloud vis-à-vis expansibility). Software layers:
  - Atmosphere interface
  - KVM
  - OpenStack
  - CentOS Linux (probably)





# Jetstream System Diagram



Dashboard

Images

Favorites

My Images

Projects

Cloud Providers

Quotas

Settings

# Search Images

Search by App Images, Tag, OS, and more

Popular Searches: [R](#) [Bisque](#) [NGS](#) [Community: Astrophysics](#)


Quick Sort:  Popularity  Recency  Rating

[Advanced Search Options](#)

Quick Filter:

View as:




## Popular Images from All Communities


 ☆

**Math Kernel Library**

[blas](#) [fft](#) [fortran](#) [lapack](#)

Community: Mathematics

 52  0  7




 ★


**RNASeq Analysis Tools**

[bowtie2](#) [blast](#) [blat](#) [edgeR](#)

[R](#) [rnaseq](#) [tophat2](#)

Community: Biology




 30  2  4


 ☆

**Atmospheric Dispersion Modeling**

[aermod](#) [aermet](#) [aermap](#)

Community: Atmospheric Sciences

 20  0  0




 ☆


**MrBayes with TreeMix**


[bayesian inference](#) [mrbayes](#)


[treemix](#)

Community: Phylogenetics

 25  1  10

 ★

 ☆

 ☆

 ☆



# Science Domains and Users

- Biology
- Earth Science/Polar Science
- Field Station Research
- Geographical Information Systems
- Network Science
- Observational Astronomy
- Social Sciences
- Jetstream will be particularly focused on researchers working in the “long tail” of science with born-digital data
- Enabling analysis of field-collected empirical data on the impact and effects of global climate change will be one of the specific foci of Jetstream
- Whatever *you* do



# 21<sup>st</sup>-century workforce development

- Jetstream will include virtual Linux desktops and applications specifically aimed to enable research and research education at small colleges and universities including HBCUs (Historically Black Colleges and Universities), MSIs (Minority Serving Institutions), Tribal colleges, and higher-ed institutions in EPSCoR States.
- Jetstream will also support deployment of user-friendly Science Gateways.



# Jetstream Deployment Partner Organizations

A seasoned team of organizations and experts:

- University of Texas Austin (TACC)
- University of Chicago (Argonne National Lab)  
(The above trio should look familiar)
- University of Arizona
- University of Texas at San Antonio (Open Cloud Lab)
- Johns Hopkins University
- Penn State University



# Globus Online – the biggest thing in big data movement



Products ▾

News ▾

About ▾

Support ▾

Lo



Research data  
management  
simplified.

79,658,453,221 MB  
TRANSFERRED



Award #1445604



[pti.iu.edu/jetstream](http://pti.iu.edu/jetstream)

# Citation

Please cite as: Stewart, C.A. 2015. Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online. Presented to Microsoft, Inc. visiting group. Indiana University, Bloomington IN, 24 Feb 2015, <http://hdl.handle.net/2022/19685>.



# License Terms

- Stewart, C.A. 2015. Big Data, Big Red II, Data Capacitor II, Wrangler, Jetstream, and Globus Online. Presented to Microsoft, Inc. visiting group. Indiana University, Bloomington IN, 24 Feb 2015. <http://hdl.handle.net/2022/19685>
- 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.
- **FOR JETSTREAM RELATED WORK:**
- This research was supported in part by the National Science Foundation through Award ACI-1445604. 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.

