Cyberinfrastructure for Research: New Trends and Tools (Part 2 of 2)

Craig Stewart

ORCID ID 0000-0003-2423-9019

Jetstream Principal Investigator

Executive Director, Indiana University Pervasive Technology Institute

30 September 2015

Presented at University of Vermont, Burlington VT

XSEDE

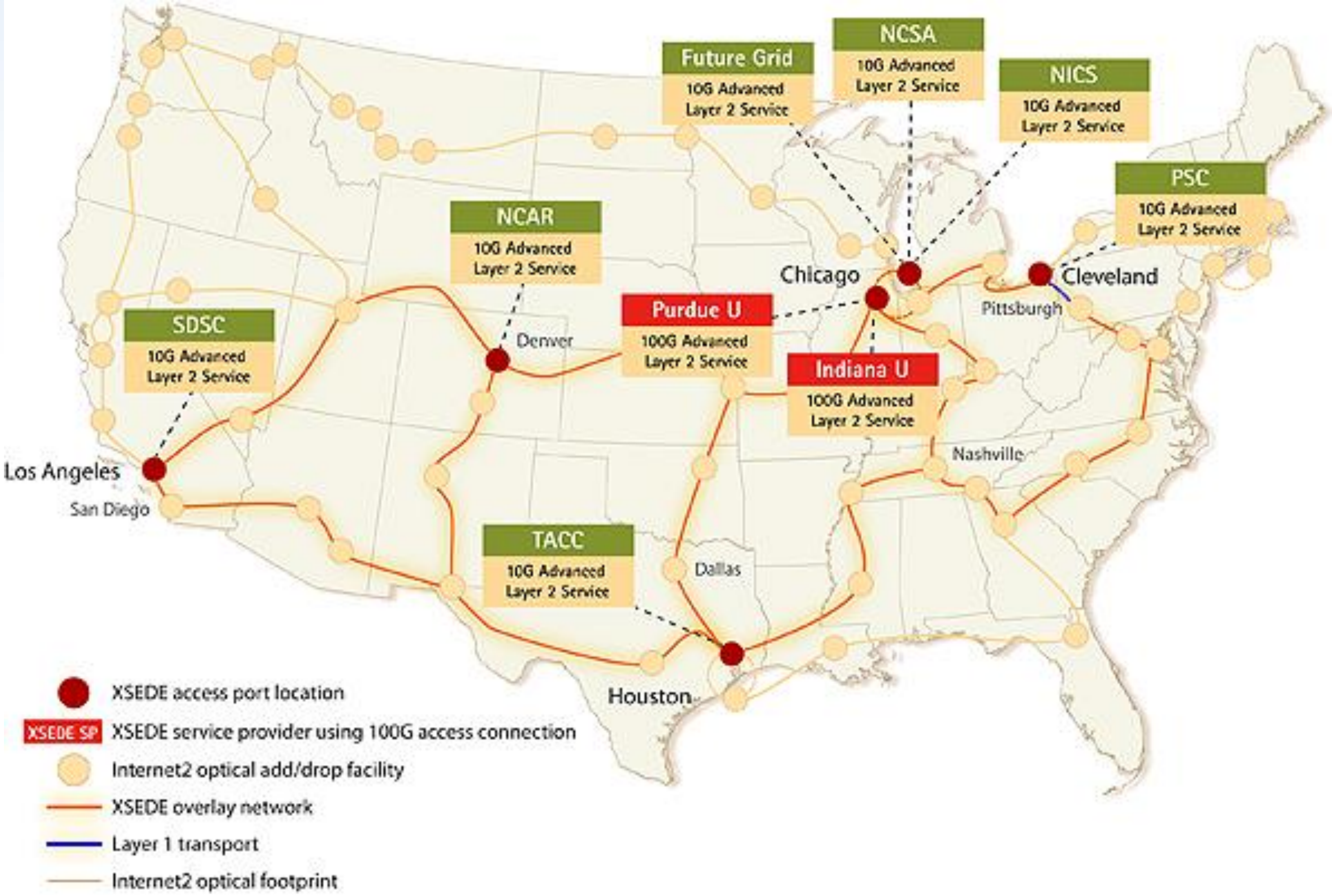
Extreme Science and Engineering
Discovery Environment

NSF

XSEDE (xsede.org) is a national source of cyberinfrastructure resources

- Allocated
  - Cycles
  - Data storage
  - Support
  - Get help the first time you apply - help@xsede.org and/or via your local campus champion
- Available to all (without allocations)
  - Globus Transfer
  - Training & curriculum materials
  - Campus Bridging

XSEDE – a national cyberinfrastructure instrument


Future Grid
10G Advanced Layer 2 Service
NCSA
10G Advanced Layer 2 Service
NICS
10G Advanced Layer 2 Service
PSC
10G Advanced Layer 2 Service
NCAR
10G Advanced Layer 2 Service
SDSC
10G Advanced Layer 2 Service
Purdue U
100G Advanced Layer 2 Service
Indiana U
100G Advanced Layer 2 Service
TACC
10G Advanced Layer 2 Service
Chicago
Cleveland
Pittsburgh
Denver
Los Angeles
San Diego
Nashville
Dallas
Houston
XSEDE access port location
XSEDE SP XSEDE service provider using 100G access connection
Internet2 optical add/drop facility
XSEDE overlay network
Layer 1 transport
Internet2 optical footprint

From xsede.org

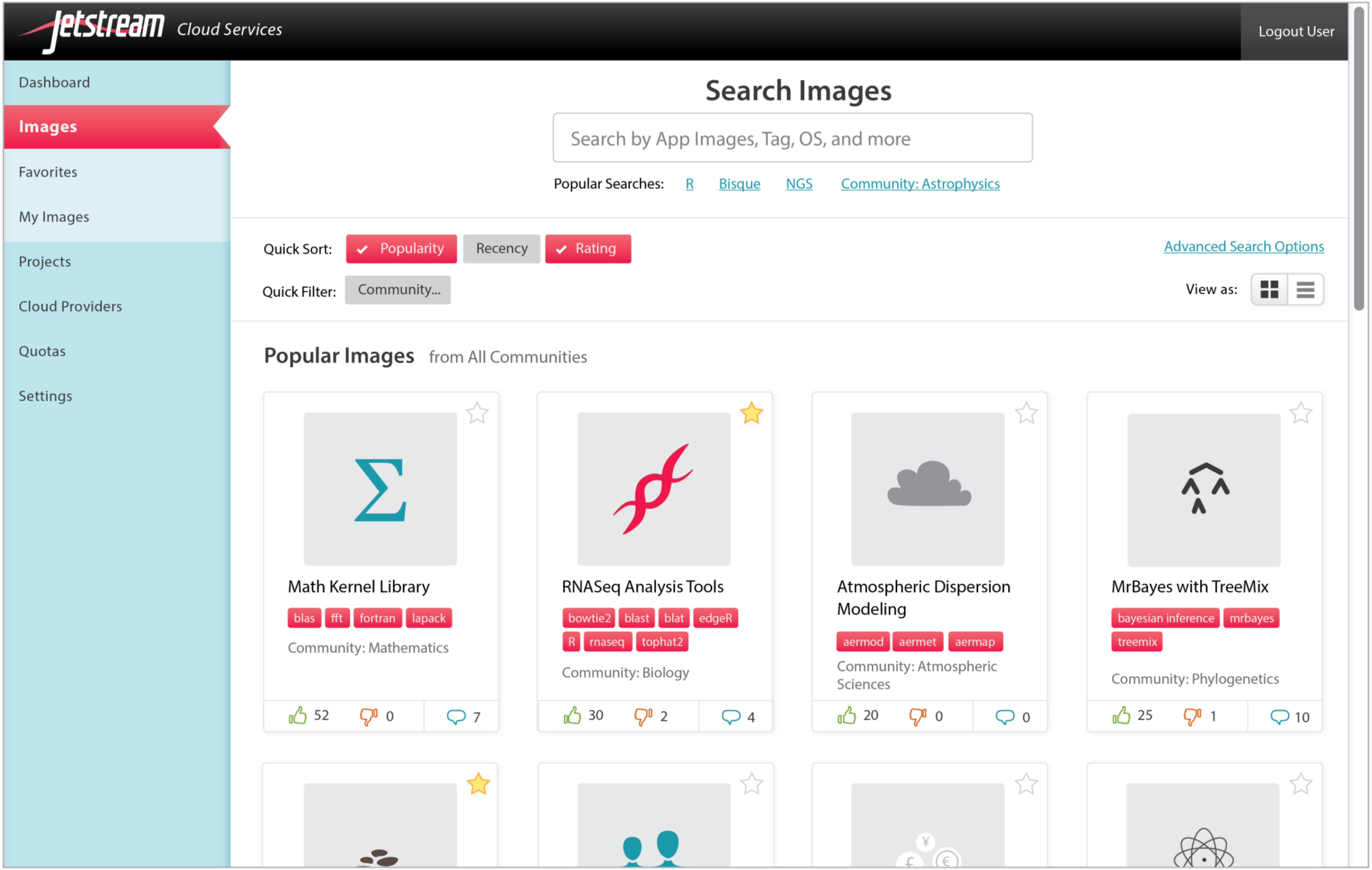
New resources to help you (focusing on easiest to use)

- Systems for you to use
  - Jetstream coming in 2016
  - Bridges coming in 2016
  - Comet available now
  - Wrangler available now
- Managing your own systems
  - XCBC (XSEDE Compatible Basic Cluster)
- Consulting Help
- XSEDE ECSS
- NCGAS (National Center for Genome Analysis Support
- All funded by federal government and available via allocations

Jetstream

A national science & engineering cloud

funded by the National Science Foundation
Award #ACI-1445604



Jetstream Cloud Services

Logout User

- 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: Community...

View as:

Popular Images from All Communities

Image	Tags	Community	👍	👎	💬
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

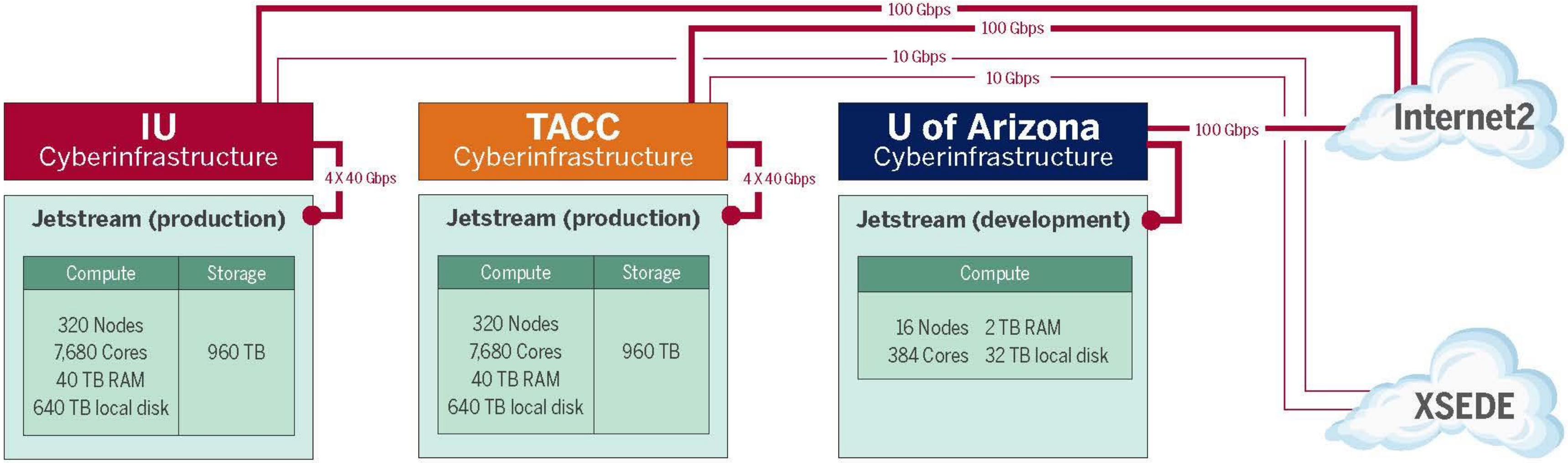
What is Jetstream?

- NSF’s first cloud for science and engineering research across all areas of NSF-supported activity.
- Jetstream will be a user-friendly cloud environment designed to give researchers and research students on-demand access to interactive computing and data analysis resources.
- Jetstream will provide a library of virtual machines from which users can select to do their research.
- Software creators and researchers will be able to create customized virtual machines or their own “private computing system” within Jetstream.
- Jetstream will enable countless discoveries across disciplines such as biology, atmospheric science, economics, network science, observational astronomy, and social sciences.
- Jetstream will support two important biology platforms: iPlant and Galaxy.

What does the name mean? 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 NSF-funded XD program and advanced cyberinfrastructure resources and users who have not 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

Jetstream System Diagram



Jetstream (production)

Compute	Storage
320 Nodes 7,680 Cores 40 TB RAM 640 TB local disk	960 TB

Jetstream (production)

Compute	Storage
320 Nodes 7,680 Cores 40 TB RAM 640 TB local disk	960 TB

Jetstream (development)

Compute	
16 Nodes 384 Cores	2 TB RAM 32 TB local disk

Science Domains and Users

- Biology
- Earth Science/Polar Science
- Field Station Research
- Geographical Information Systems
- Network Science
- Observational Astronomy
- Social Sciences
- Jetstream will focus on researchers working in the “long tail” of science with born-digital data.
- A special focus will be enabling analysis of field-collected empirical data on the impact and effects of global climate change.
- Whatever *you* do …. Unless you do large-scale parallel computing

XSEDE
Extreme Science and Engineering
Discovery Environment

BRIDGES

A PITTSBURGH SUPERCOMPUTING CENTER RESOURCE

Connecting Researchers, Data & HPC

Nick Nystrom · nystrom@psc.edu

3-Slide Summary · May 14, 2015

Bridging to Nontraditional HPC Users and Enabling HPC + Big Data Workflows

Leveraging PSC's expertise with shared memory, *Bridges* will feature 3 tiers of large, coherent shared-memory nodes – 12TB, 3TB, and 128GB – to support a uniquely flexible and user-friendly environment:

- Interactivity is the feature most frequently requested by nontraditional HPC communities and for doing data analytics and testing hypotheses.
- Gateways and tools for gateway building will provide easy-to-use access to *Bridges*' HPC and data resources, reaching large numbers of users who aren't programmers.
- Database and web server nodes will provide persistent NoSQL and relational databases to enable data management, workflows, and distributed applications.
- High-productivity programming languages & environments (R, Python, MATLAB, Java, Hadoop, etc.) will let users scale familiar applications and workflows.
- Virtualization will allow users to bring their particular environments for portability, reproducibility, and security and provide interoperability with clouds.
- Campus bridging will streamline interoperation with campus resources and enable burst offload capability through a pilot project with Temple University.

Interest from new communities is already very high: examples include the digital humanities, machine learning, statistics, genomics, and radio astronomy.

Gateways to Discovery: Cyberinfrastructure for the Long Tail of Science

ACI-1341698

What is Wrangler?

- Wrangler is a new data-intensive supercomputing system.
- Built from the ground up for data-intensive applications.
- HPC and “Big Data” have a lot in common
  - The overlap isn’t 100% in all applications.
  - Exascale computers will generate phenomenal amounts of data, but *every* data problem will map perfectly.
  - Mostly a difference in data access patterns (small random reads for data vs. large sequential writes for HPC checkpoints)
    - Centralized vs. distributed file systems (don’t try running Hadoop MapReduce on HPC hardware like Stampede)
    - Scratch file system vs. dedicated services supporting persistent data
- New technologies can bridge the shortcomings of current HPC Cluster architectures and policies.

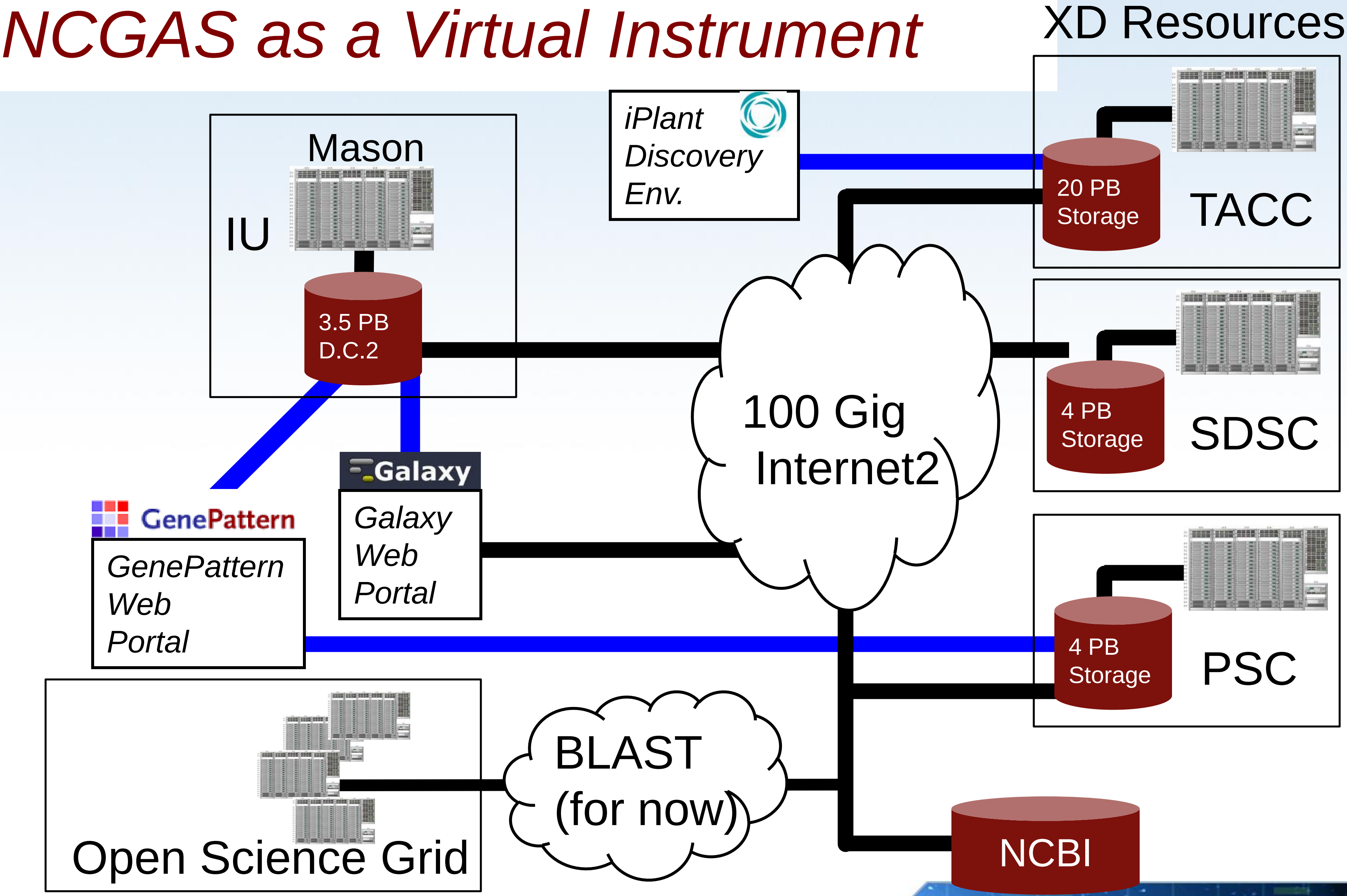
Campus Bridging – XSEDE National Integration Toolkit (XNIT)

- Software tools to:
  - Make it easier for your local systems administrators to manage your local clusters.
  - Make it easier for you to make your local clusters more consistent with systems supported by XSEDE (diversity of names and partners notwithstanding, there is a lot of consistency across systems).
  - Subscribe to the tools you want and ignore the ones you don’t
  - Build a cluster from scratch

National Center for Genome Analysis Support (NCGAS) Service Model

- Research design support
- Bioinformatics expertise
- Web workflow composers (Galaxy, GenePattern)
- Optimized software applications (esp. Trinity)
- High performance computing resources, esp. large-memory clusters = Mason
- Storage for data and dissemination of results
- Training and outreach to research community

NCGAS as a Virtual Instrument



XSEDE ECSS (Extended Collaborative Support Program)

The Extended Collaborative Support Service (ECSS)

improves XSEDE user community productivity through:

- Successful, meaningful ***collaborations***
- Well-planned ***training activities***

These:

- Optimize applications.
- Improve work and data flows.
- Increase effective use of the XSEDE digital infrastructure.
- ***Broadly expand the XSEDE user base by engaging members of under-represented communities and domain areas*.**

ECSS Major Accomplishments

- Significantly increased user productivity and user capability
  - e.g. median code speedup 2.25x, highest speedup 126x, over 200 live training/outreach events in PY3
- Expertise available in many fields
  - over 50 expertise areas
- Sometimes serve as an intellectual commons bringing disparate research groups together for increased productivity
  - e.g. among users running large-scale genomics calculations

But you do have to apply for resources

- *Resources are available for use in research projects by faculty, staff, and students and to support classroom education.*
- Go to xsede.org and make a portal account (easy)
- For resources allocated through XSEDE (Comet, Wrangler now; ECSS support now; Mason time now) fill out application form at https://www.xsede.org/allocations. Start with a startup allocation!
- Help from
  - help@xsede.org
  - campusbridging@xsede.org
  - jethelp@iu.edu
  - ncgas@iu.edu
- Ask for help asking for help!

You do NOT need current NSF funding to use XSEDE resources!

- If you have current funding from a federal funding agency, your work is assumed to have been (positively) peer reviewed. Your proposal review will look at appropriateness of the resources you request relative to your research and to priority within available resources.
- If you do not have current funding, your review will include a review of your research and the cyberinfrastructure resources you request.
- Review criteria for startup (initial small) allocations are liberal, erring on the side of granting people access. The same goes for requests for resources supporting educational activities.
- Like any NSF-funded project, XSEDE aims to have important broader impacts. Support for researchers in an EPSCoR State is a broader impact. (So those from Kentucky have a factor in their favor.)

This is an ecosystem issue

- National Strategic Computing Initiative
- XSEDE
- Campus
  - Develop a diverse user base, diverse needs.
  - Emphasize local strengths in science, humanities, and arts.
  - Local strategy and consistency is essential (You need today’s Publius Cornelius Scipio, not today’s Hannibal.)
  - Work like #$%#$% to get federal monies, as OPM is the best.
  - Foster a local community and invest in support first, and hardware second, and at a level you maintain. No moonshots. Have sufficient local resource as an onramp to the national resources.
  - Faculty and staff who believe in the common goal of the university need to value each other and demonstrate that in collaboration.

But it will never be perfect - We Live the Myth of Sisyphus

"*The struggle itself...is enough to fill a [person's] heart. One must imagine Sisyphus happy*."
–Albert Camus

Sisyphus (1548-1549) by Titian, Prado Museum, Madrid, Spain
http://en.wikipedia.org/wiki/File:Punishment_sisyph.jpg
This work is in the public domain in the United States, and those countries with a copyright term of life of the author plus 100 years or fewer.

Jetstream Collaborators

- University of Chicago - Globus
- Arizona University – iPlant
- Johns Hopkins University and Penn State University
- Cornell University –Ms. Susan Mehringer, Lead. Cornell® Virtual Workshops about Jetstream and applications running on Jetstream.
- University of Arkansas at Pine Bluff – Dr. Jesse Walker, lead. Cybersecurity education, Minority Serving Education outreach.
- University of Hawaii – Dr. Gwen Jacobs, lead. EPSCoR early adopter/user. Jacobs will chair Science Advisory Board.
- National Snow and Ice Data Center (NSIDC) – Dr. Ron Weaver, lead. Data retrieval from NSIDC, application integration with ice-sheet analysis applications.
- University of North Carolina, Odum Center –Dr. Thomas Carsey , lead. Data retrieval from Dataverse Network.
- National Center for Genome Analysis at Indiana University, providing genome analysis software. Includes TACC, PSC, and SDSC as partners.

UC San Diego

SDSC
SAN DIEGO SUPERCOMPUTER CENTER

DELL

intel

INDIANA UNIVERSITY

AEON COMPUTING

Mellanox TECHNOLOGIES

This work supported by the National Science Foundation, award ACI-1341698.

NCGAS Partners

NIH

National Institutes of Health

NSF

TACC

TEXAS ADVANCED COMPUTING CENTER

Powering Discoveries That Change The World

PITTSBURGH SUPERCOMPUTING CENTER

XSEDE

Extreme Science and Engineering Discovery Environment

BROAD INSTITUTE

Open Science Grid

SDSC

SAN DIEGO SUPERCOMPUTER CENTER

Acknowledgments & Disclaimers

- Thanks to Nick Nystrom of the Pittsburgh Supercomputing Center for slides about the new Bridges System. Bridges is supported by NSF award 1445606.
- Thanks to Richard Moore of the San Diego Supercomputer Center for slides about Comet. Comet is supported by NSF award 1341698.
- Thanks to Daniel Stanzione of the Texas Advanced Computing Center for slides about Wrangler. Wrangler is supported by NSF award 1341711.
- Jetstream is supported by NSF award 1445604 (Craig Stewart, PI).
- XSEDE is supported by NSF award 1053575 (John Towns, UIUC, PI).
- This work was also supported by the Indiana University Pervasive Technology Institute, which was initiated with major funding 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 National Science Foundation (NSF) or other supporting organizations.

Questions?????

License Terms

- Please cite as Stewart, C.A. 2015. Cyberinfrastructure for Research: New Trends and Tools (Part 2 of 2). Presentation. University of Vermont, Burlington, VT. 30 September 2015.
- 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.
- Jetstream 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.

Our reach will forever
exceed our grasp, but,
in stretching our horizon,
we forever improve our world.

XSEDE

Extreme Science and Engineering
Discovery Environment