

NCGAS: Providing National Cyberinfrastructure to Biologists, esp. Genomicists.

Thomas G. Doak, PI and Manager
National Center for Genome Analysis Support



INDIANA UNIVERSITY

An outline:

- The science and research NCGAS addresses:
 - as an NSF service (our own grant)
 - beyond NSF (*i.e.* ITCR) (on others grants)
- What tools and infrastructure XSEDE provides to researchers
 - e.g. Jetstream
- Docker (shifter)



NCGAS's primary goals:

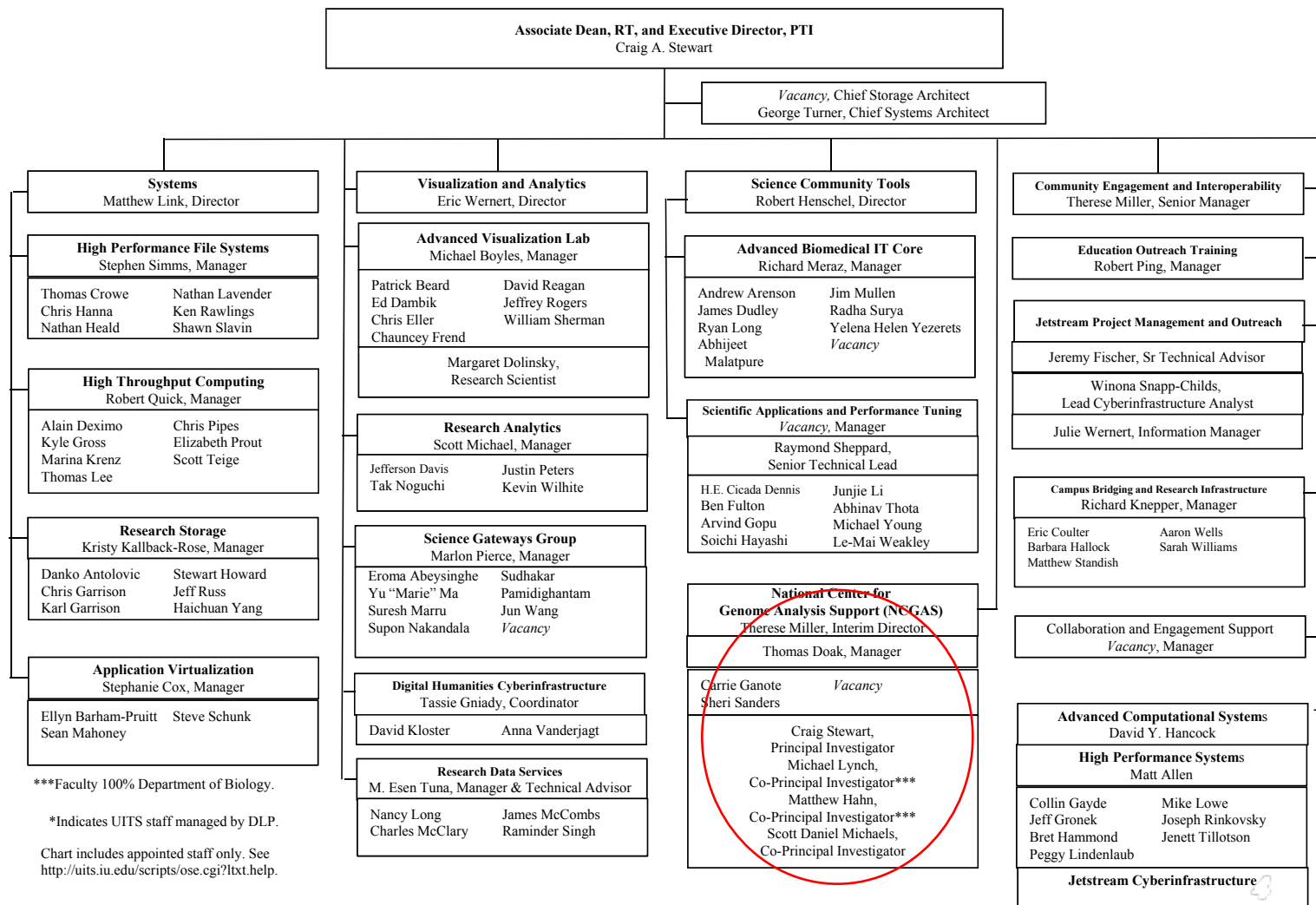
- Provide bioinformatics expertise
- Maintain a curated set applications
- Provide access to HPC resources, esp. large-memory clusters = Mason, Bridges
- Build Galaxy instances for our software
- Pursue outreach to biologists

NCGAS is embedded in Research Technologies

Indiana University

Research Technologies & PTI Service & Cyberinfrastructure Centers

September 2016

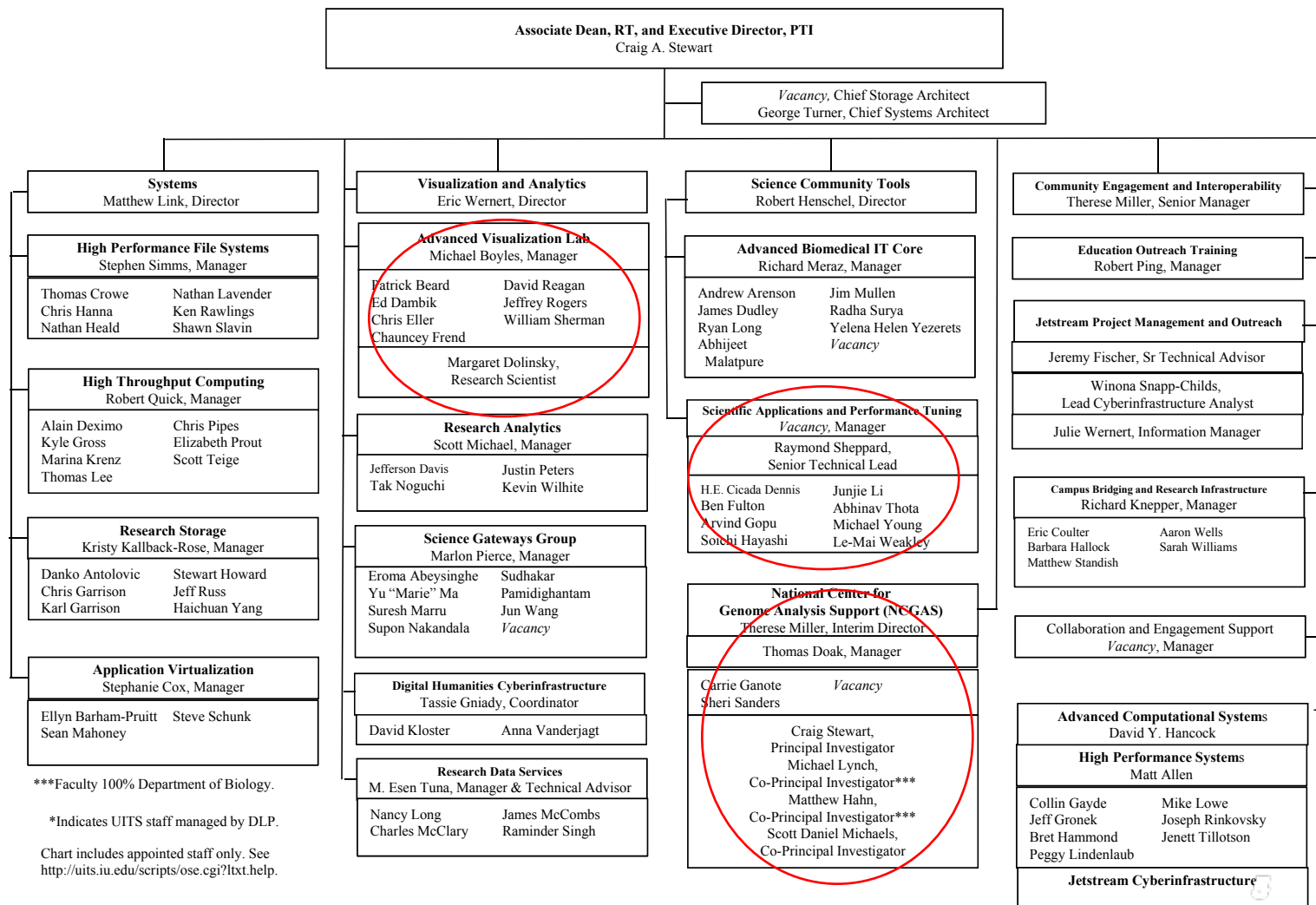


NCGAS is embedded in Research Technologies

Indiana University

Research Technologies & PTI Service & Cyberinfrastructure Centers

September 2016





Supporting NCGAS Genomics Research at PSC

Philip D. Blood, Ph.D.
Senior Computational Scientist
Pittsburgh Supercomputing Center

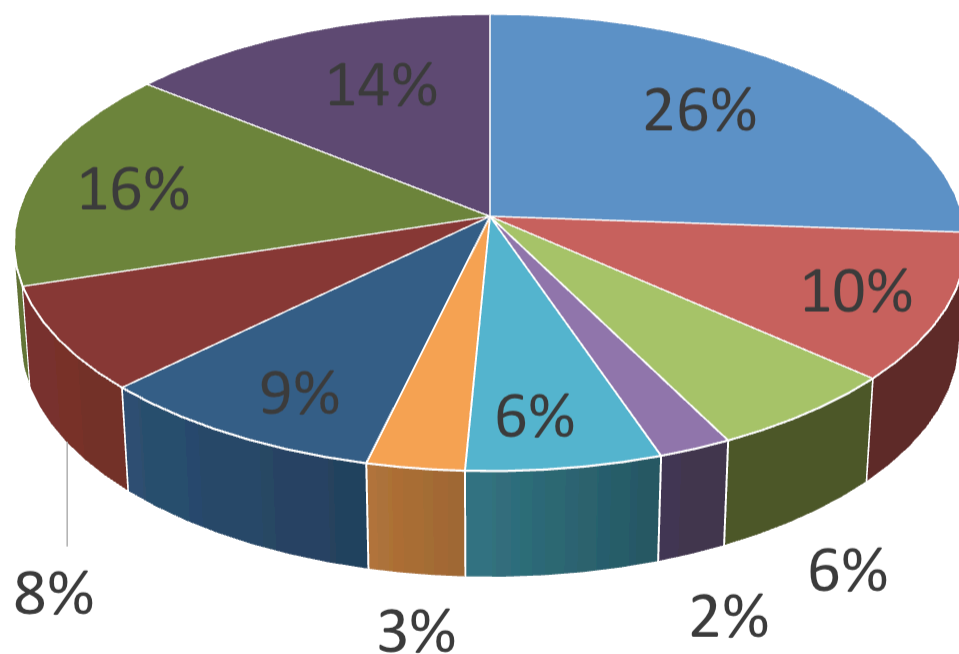
Bridges:

- 2×16 TB of cache-coherent shared memory, 4096 cores
- ideal for genome sequence assembly
- High bandwidth, low latency interprocessor communication

*Bridges leverages its large memory for **interactivity** and to seamlessly support applications through **virtualization**, **gateways**, familiar and **productive programming environments**, and **data-driven workflows**.*

From our recent NSF survey:

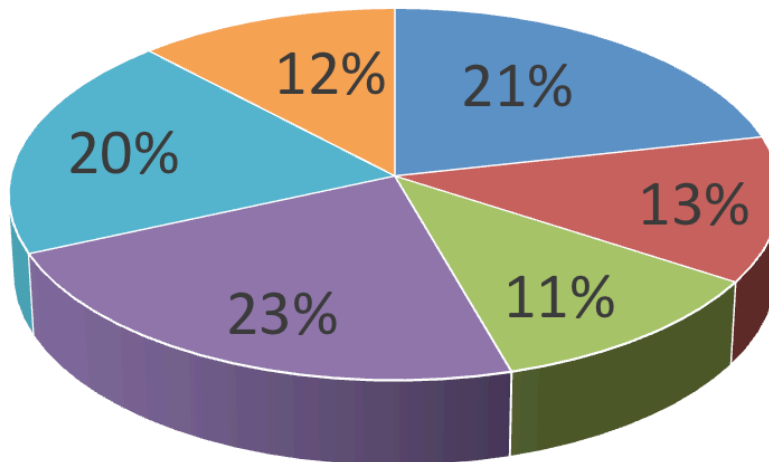
Types of Research (Multi-Pick)



- RNA-Seq approaches?
- Microbiome sequencing?
- Whole genome structural analysis, such as nucleosome mapping or high-resolution DNase sensitivity?
- Gene function analysis using parallel transposon insertion sequencing?
- Protein-DNA interactions with ChIP-seq?
- Methylation studies with whole-genome bisulfate sequencing?
- Gene expression profiles with microarrays?
- Proteomics and Mass Spec?
- Genome annotation and ortholog discovery?

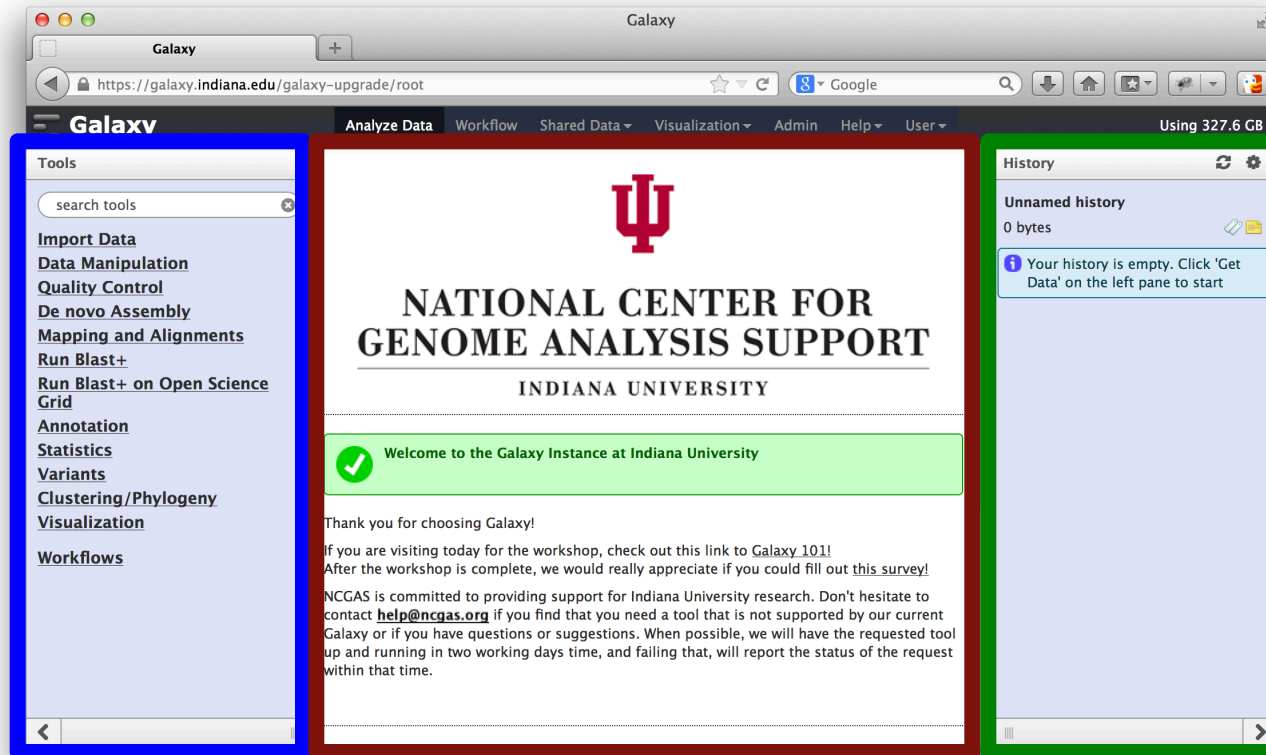
From our recent NSF survey:

Would any of the follow services
be helpful to you (Multi-Pick)



- Bioinformaticians on-call.
- Large RAM infrastructure.
- Fast CPU nodes.
- Installed curated genomics applications Installation of published bioinformatic applications.
- Mounting of applications as web-accessible Galaxy tools.
- Rapid access to reference data for analysis

Galaxy Anatomy and Physiology



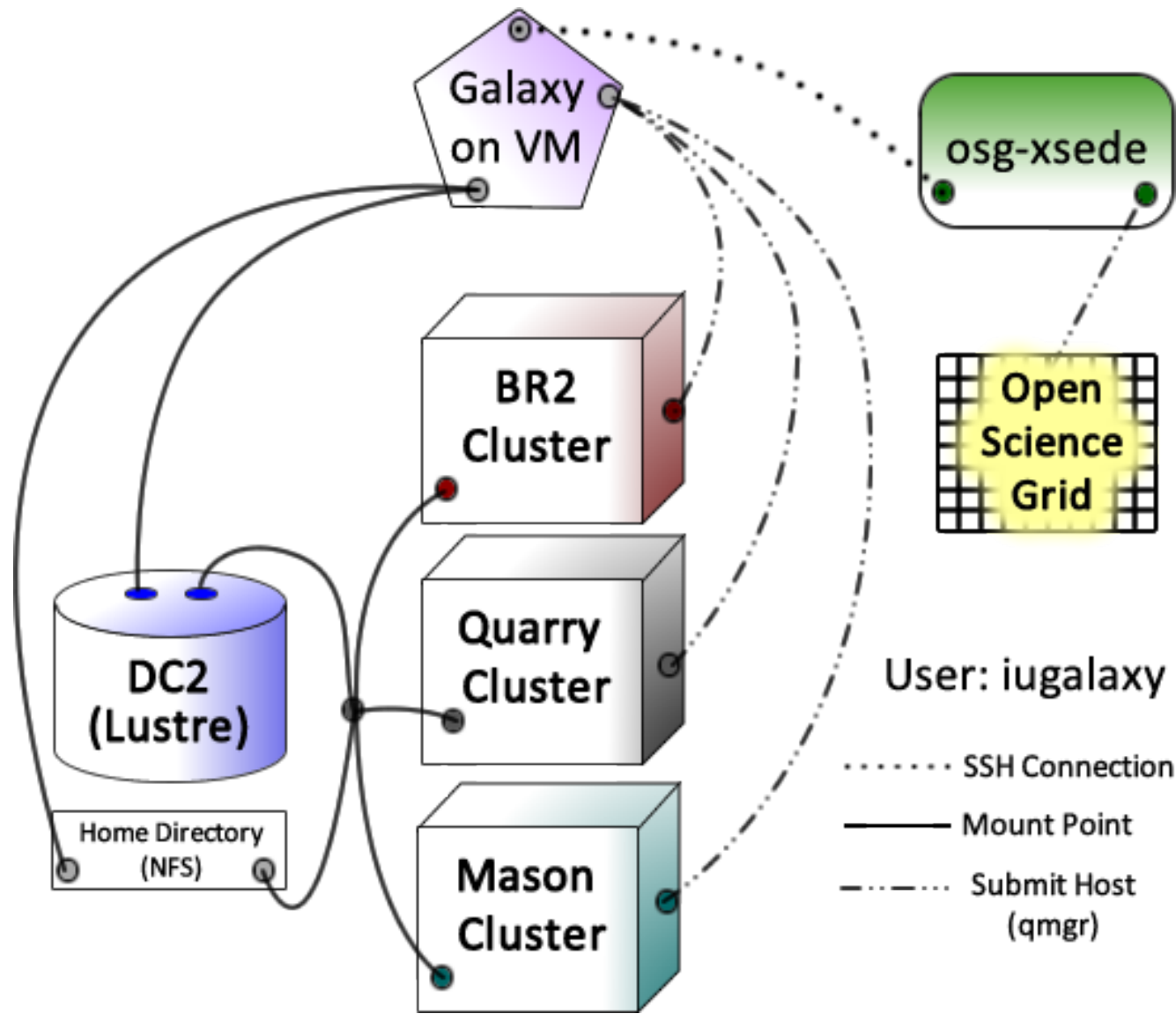
Tool bar – contains the available steps to apply to data

History – shows steps previously taken to manipulate input data sets

Focus pane – shows options, parameters, and output for current item.

National Center for Genome Analysis Support: <http://ncgas.org>

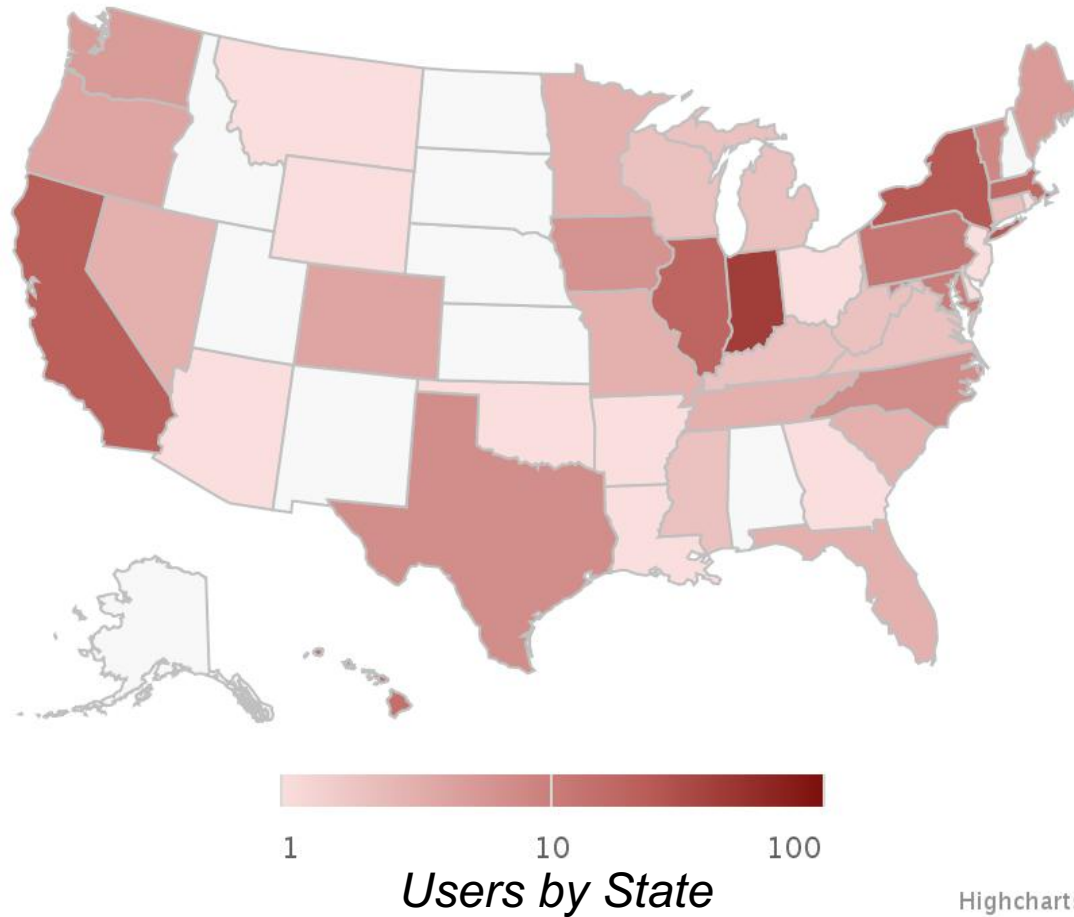
GALAXY.NCGAS.ORG Model



We're reaching most states

National Center for Genome Analysis Support Users 2016

Representing approximately 105 institutions



Environmental Genomics Workshop 2016

MDIBL hosted Environmental Genomics in Salisbury Cove, Maine. Over one week, 9 students and 9 post doc/faculty from 14 universities learned to design RNA-seq experiments, create *Daphnia* RNA-seq libraries, manage and analyze the data, and present their results.

NCGAS partnered with the workshop for the first time in 2016 and provided:

- Three reserved Karst nodes, totaling
 - 48 processors
 - 48 GBs of RAM
- On site training and consultation on cluster use, bioinformatics, and statistics. Immediate coordination with IU sysadmin.
- Rapid processing of ~250Gb of sequence data, when data arrival was delayed

As a result of partnership with NCGAS, the workshop was able to include 4.5x more data throughput, enabling more realistic and complex experimental designs.





NATIONAL CENTER FOR GENOME ANALYSIS SUPPORT

INDIANA UNIVERSITY

The fine print

We ask that you
acknowledge our grant in
any published work that
uses our resources.
Collaborations and
authorship are requested
for intellectual
contributions.

THE FACTS

- 16-nodes, 500GB RAM
- 10TB project space
- Bioinformatics software
- Galaxy instance
- 50TB archive space/user



Links from NCGAS page



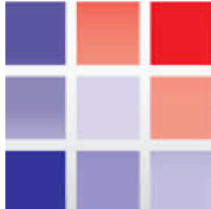
IU students and faculty have access to Galaxy @ IU using their Indiana University credentials. Even with access to the resources available at IU, the interface between users and computer clusters can be daunting. We provide support to IU affiliates through Galaxy to accomplish their bioinformatics analyses without the need for a degree in computer science.

[Click here to enter Galaxy @ IU](#)



Trinity CTAT Galaxy, hosted by Indiana University and the Broad Institute, is a free-to-use public interface for Trinity users

[Click here to enter Trinity_CTAT Galaxy](#)



GenePattern is a freely available computational biology open-source software package developed at the Broad Institute of MIT and Harvard, for the analysis of genomic data. NCGAS now hosts a free GenePattern server, with increased computational resources.

[Click here to enter IU's Public GenePattern server](#)



Galaxy Main, hosted by Penn State University and Emory University, is a free-to-use public service that includes hundreds of tools and a server with 250 GB of storage space per user. Our own Galaxy instances are based off of this technology.

[Click here to enter Galaxy Main](#)

Trinity Galaxy Home Page @ IU

Galaxy

Analyze DataWorkflowShared Data▼Visualization▼AdminHelp▼User▼

Using 4.7 GB

Tools

search tools

Import Data

De novo Assembly

Trinity De novo assembly of RNA-Seq data using Trinity on the Karst cluster


Trinity De novo assembly of RNA-Seq data using Trinity on the Karst cluster


Helper Tools

Fasta Tools

Workflows

■ [All workflows](#)


**National Center for
Genome Analysis Support**
Indiana University Pervasive Technology Institute

 **Welcome to the Trinity Galaxy Instance**

Thank you for choosing Galaxy!

Get started with some help [moving files into Galaxy](#). Feel free to visit our [FAQ page](#) for additional information.

We are committed to helping you succeed with your research. Don't hesitate to contact help@ncgas.org if you need help or if you have questions or suggestions.

 This Galaxy instance is running on hardware that is scheduled to be unavailable on the first Tuesday of every month for maintenance. Jobs that are started before this time will resume after maintenance.

The Trinity project is supported by the National Cancer Institute of the National Institutes of Health under award number U24CA180922.

This instance of the Galaxy is installed and maintained by National Center for Genome Analysis Support [NCGAS](#) (NSF Award #1062432)

History

search datasets

Unnamed history
127 shown, 27 [deleted](#)
3.9 GB

154: Trinity on data 1 and data 2: Assembled Transcripts

153: Trinity on data 1 and data 2: log

152: Trinity on data 1 and data 2: Assembled Transcripts

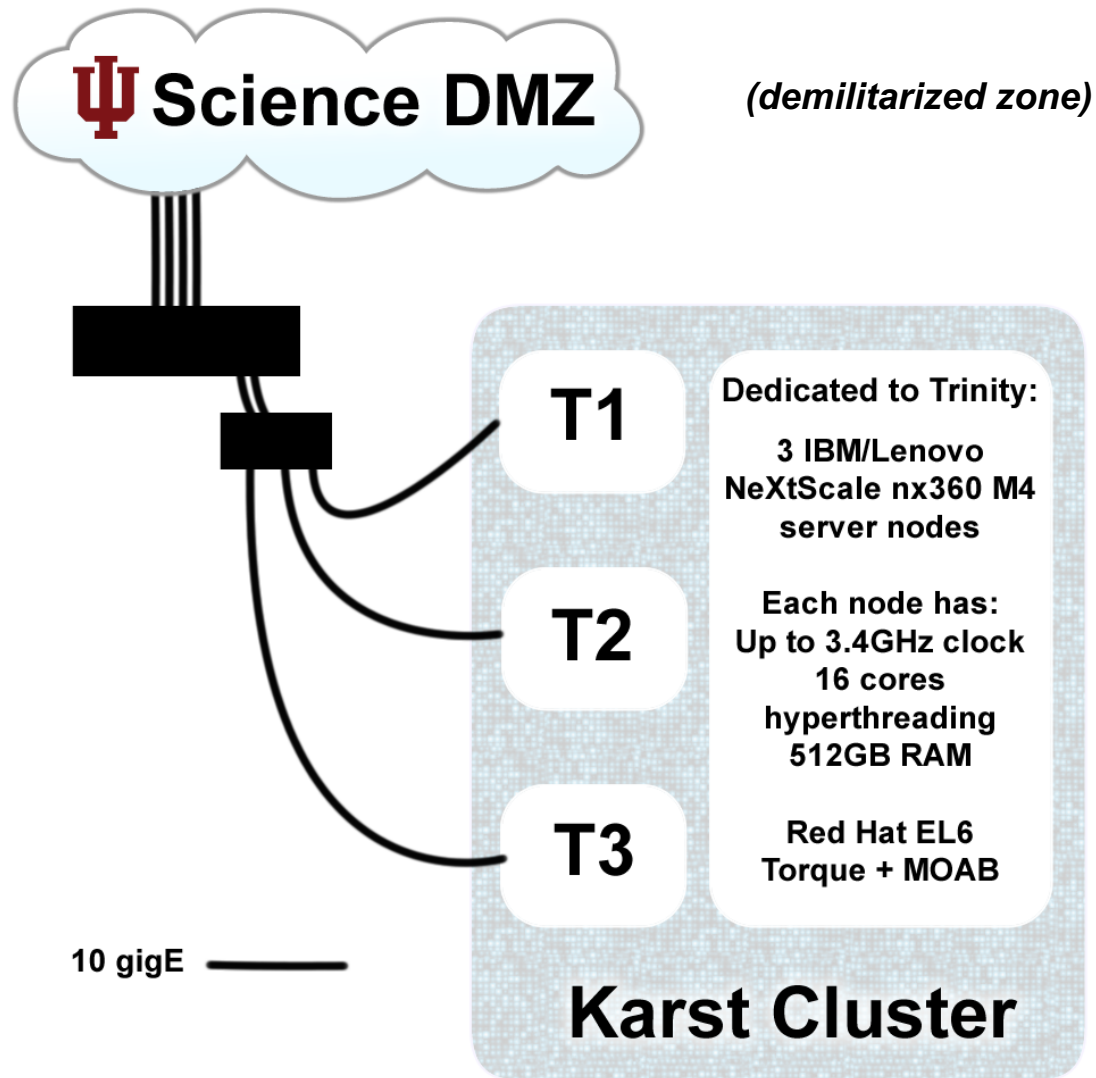
151: Trinity on data 1 and data 2: log

148: Trinity on data 1 and data 2: Assembled Transcripts

147: Trinity on data 1 and data 2: log

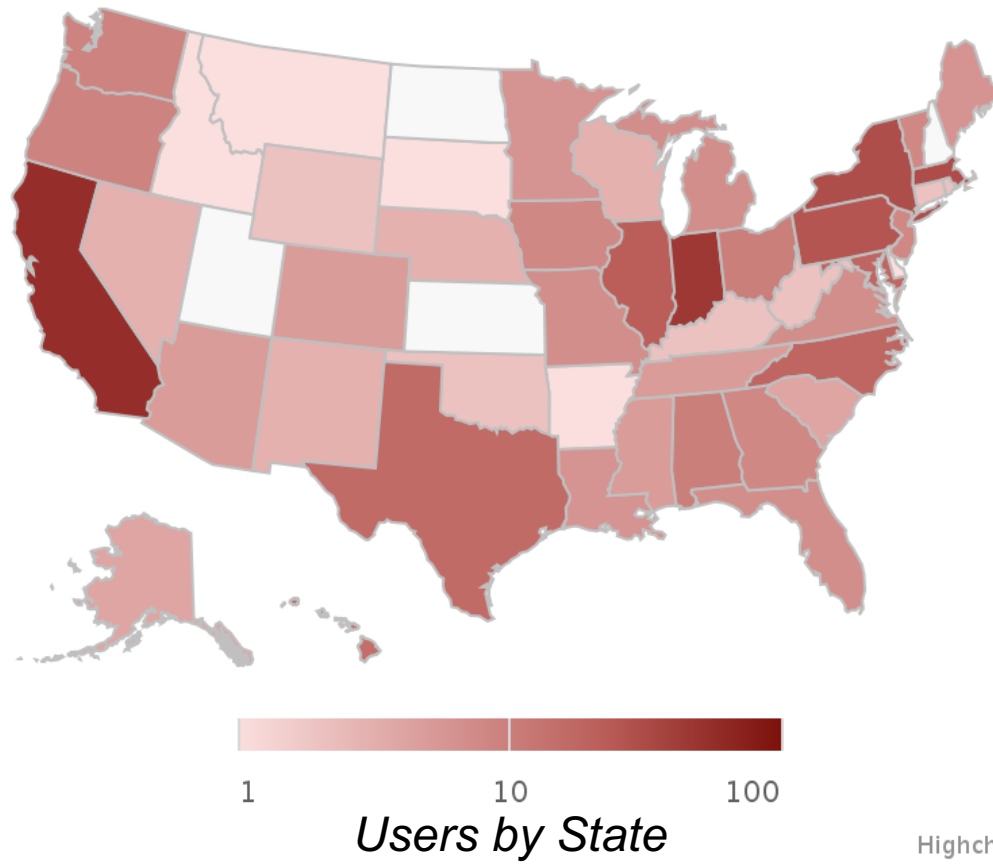
146: Trinity on data 1 and data 2: Assembled Transcripts

The Trinity Grant punched Karst “condo” nodes



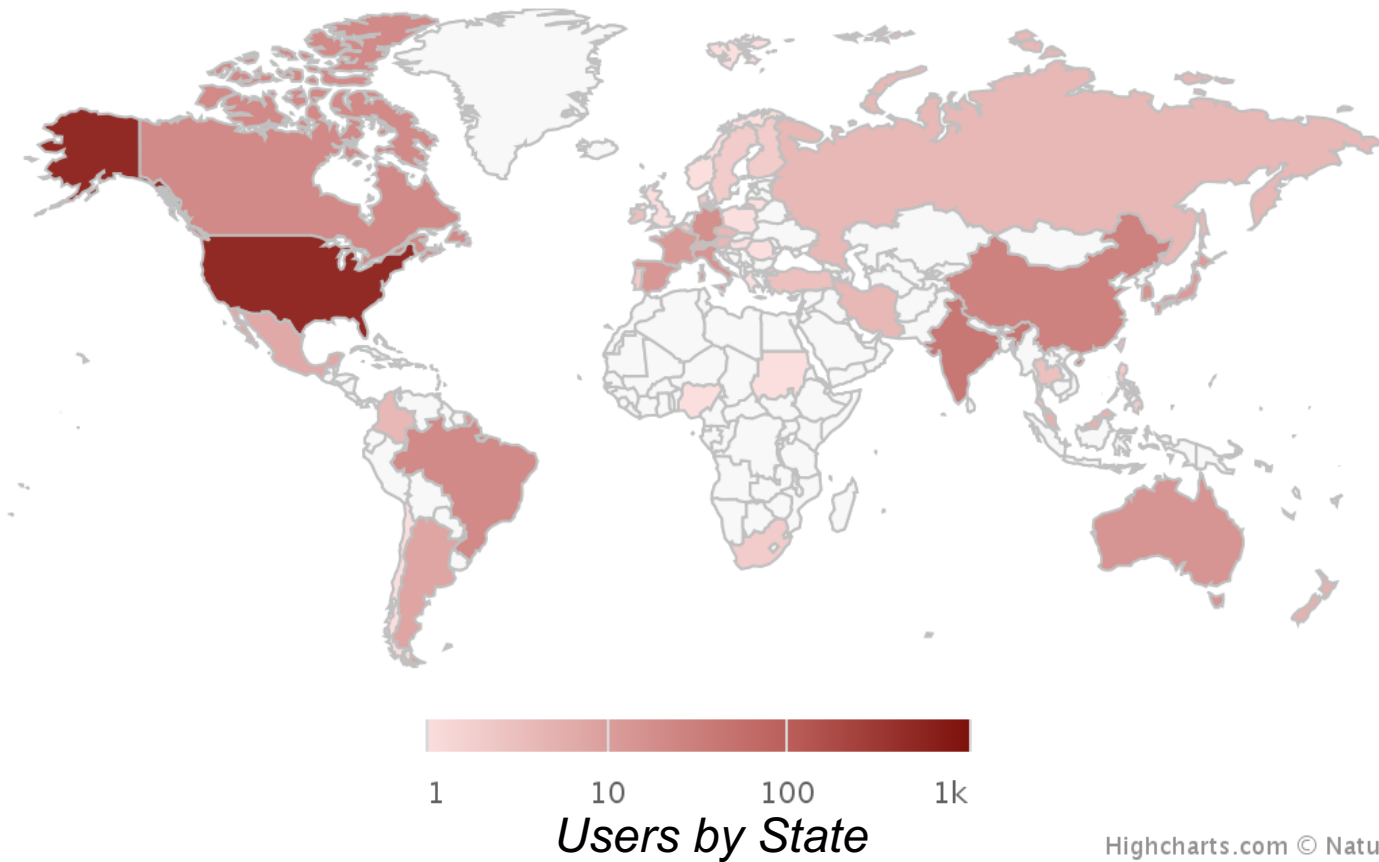
National Center for Genome Analysis Support and Trinity Galaxy Users 2016

Representing approximately 233 institutions



National Center for Genome Analysis Support and Trinity Galaxy Users 2016

Representing approximately 545 institutions in 49 countries

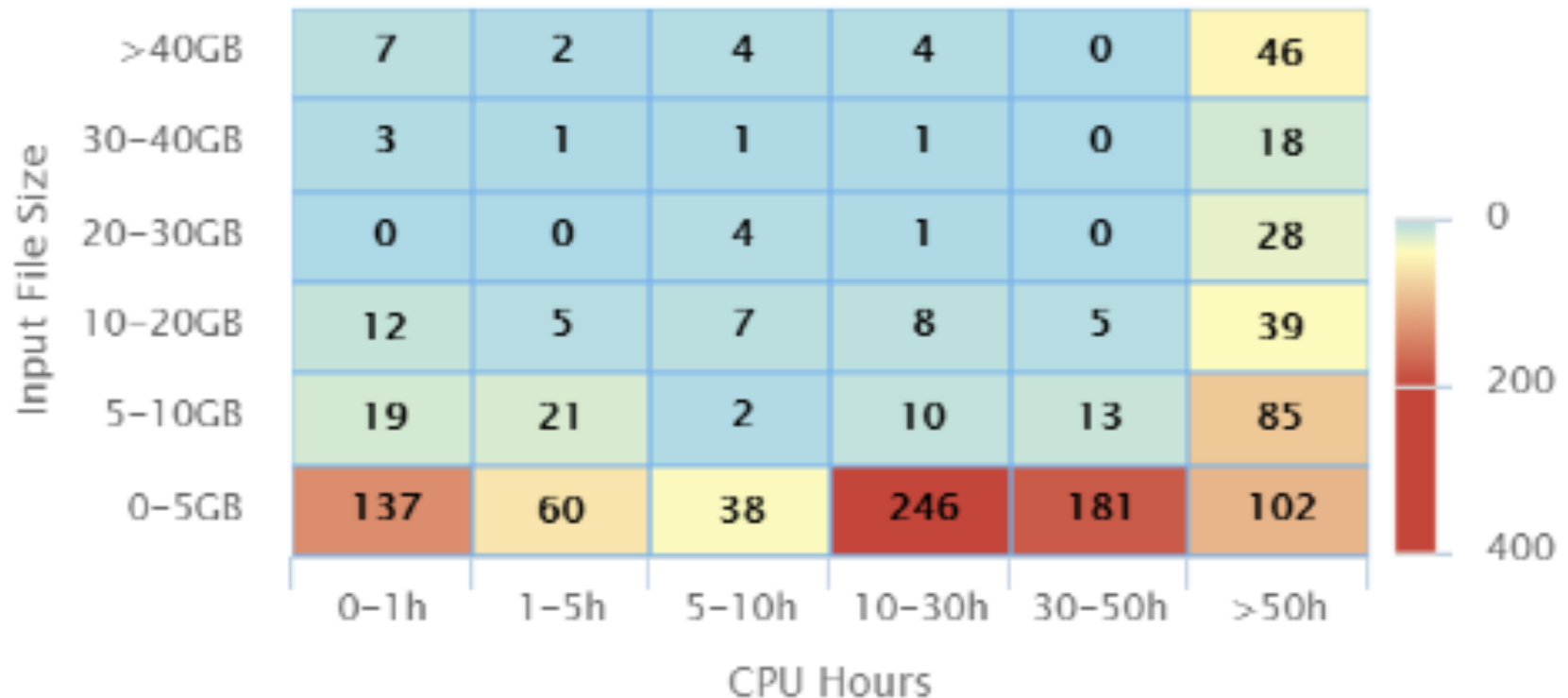


Trinity metrics: speed, use, memory, etc.

Galaxy Job CPU Hours by Input File Size



Data from 2014-04-03 to 2016-09-02

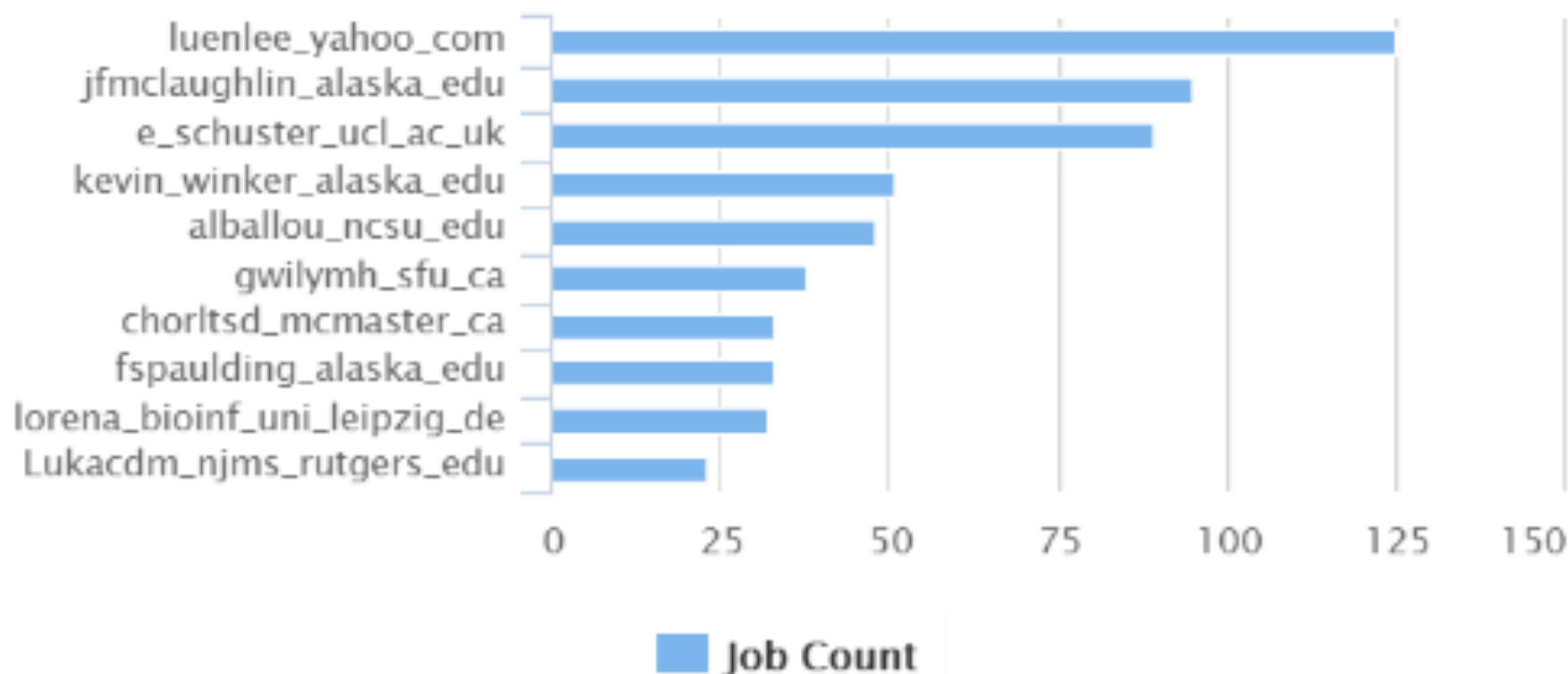


Trinity metrics: Just for fun

Top Ten Galaxy Users by Job Submissions



Data from 2014-04-03 to 2016-09-02



Use GenePattern

The GenePattern team and collaborating organizations maintain several servers that can be used without installing any software. All that is needed is to register (Note that each server must be registered for separately). More information about the servers is provided below.

GenePattern @ Broad

<https://genepattern.broadinstitute.org/gp>

The Broad Institute hosts a publicly available GenePattern server.



- The job purge for the public Broad-hosted server is set to 7 days.
- There is a 30 GB quota on jobs and uploaded data.
- Most of the modules and pipelines available from the Broad Institute (see the [Modules](#) page of the GenePattern web site) are available on the Broad-hosted server. Several modules are available *only* on the Broad-hosted server because they require customized server configuration. (If you are interested in these modules, [contact us](#) for more information). A small number of modules are not available on the Broad-hosted server because they run only on the Windows platform; the Broad-hosted server runs under Unix.

GenePattern @ Indiana University

<http://gp.indiana.edu/gp/>

The GenePattern team in collaboration with Indiana University's (IU) National Center for Genome Analysis Support (NCGAS) hosts a public server on IU's high performance computing system. This server has more capacity to better accommodate next generation sequencing analysis and other compute intensive analyses.



- **This resource is only available for academic and non-profit users.**
- The job purge for the IU server is currently set to 30 days, as opposed to 7 days on the Broad server, so your jobs will remain on the server longer.
- There is currently no quota on the IU server, with regard to jobs or data upload. This may change in the future.
- The server is unavailable on the first Tuesday of every month for planned maintenance. We will communicate any other planned downtime as far in advance as possible.

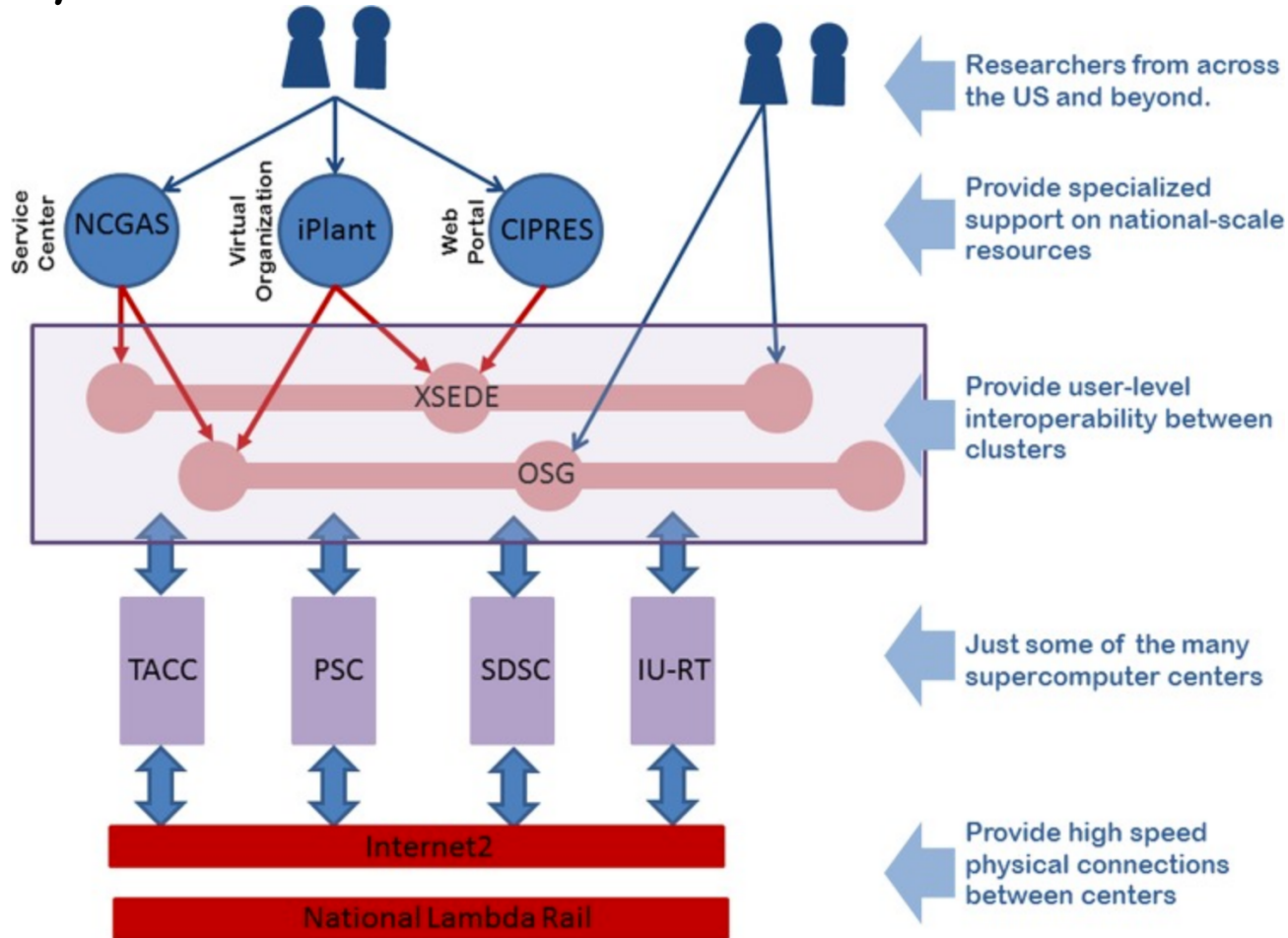
@Pittsburgh (PSC) too

A horizontal banner with a dark blue background. On the left, there are several glowing, textured spheres of varying sizes, some connected by thin white lines. In the center, a large, translucent blue sphere with a bright light source inside is visible. On the right, the text 'XSEDE' is written in large, bold, white capital letters. Below it, the text 'Extreme Science and Engineering Discovery Environment' is written in a smaller, white, sans-serif font. The overall aesthetic is futuristic and scientific.

XSEDE

Extreme Science and Engineering
Discovery Environment

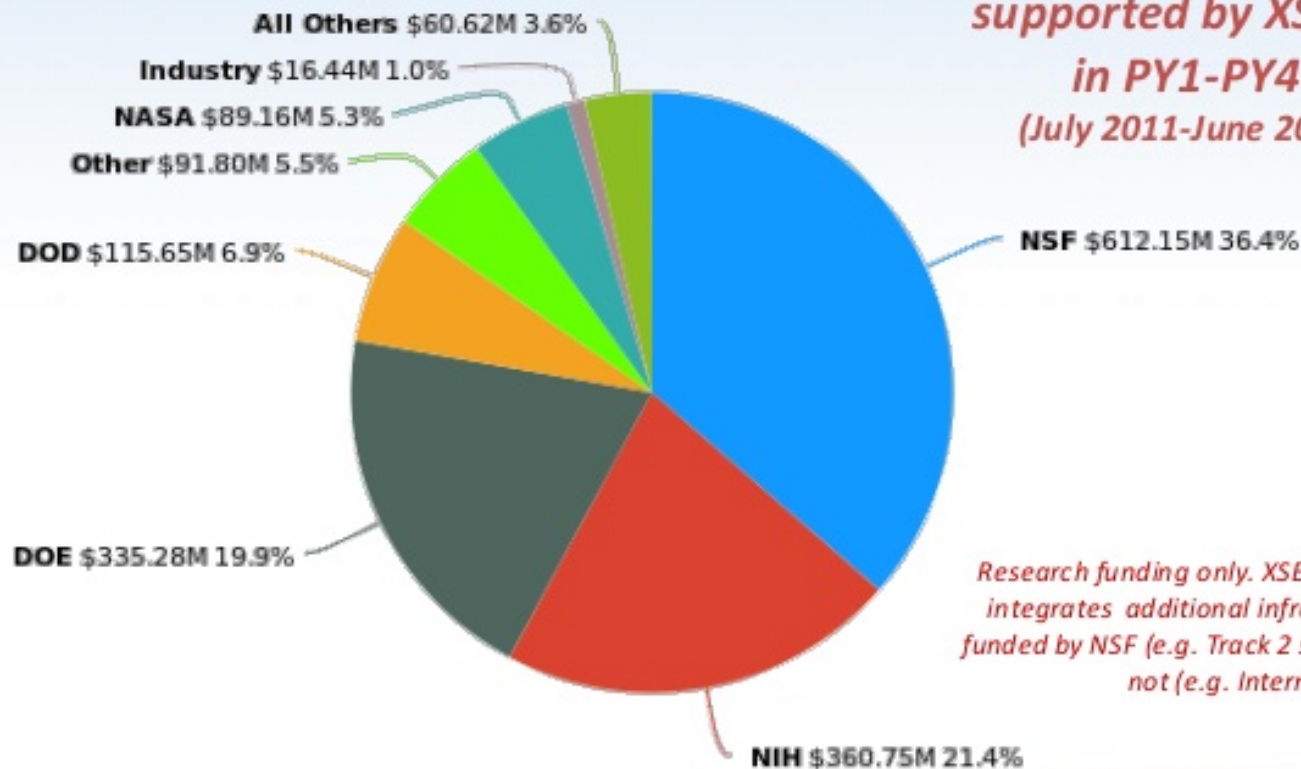
***A Roadmap to the Research Information Superhighway:
over 200 supercomputer centers are interconnected across a series of high
speed physical networks. The resources in these centers are shared across
organizations such as XSEDE and OSG. Specialized centers use XSEDE and OSG
to support specialized user communities.***



Anyone can use XSEDE

Total Research Funding Supported by XSEDE in Program Years 1-4

*\$1.68 billion in research
supported by XSEDE
in PY1-PY4
(July 2011-June 2015)*



Research funding only. XSEDE leverages and integrates additional infrastructure, some funded by NSF (e.g. Track 2 systems) and some not (e.g. Internet2).





Research Computing on Cloudy Platforms

Jetstream: A national research and educational cloud

J. Michael Lowe (jomlowe@iu.edu)

Jetstream System Engineer
IU High Performance Systems

George Turner (turnerg@iu.edu)

Chief Systems Architect
IU Research Technologies

Operating Innovative Networks Workshop, Indiana University - Bloomington, 12-July-2016



funded by the National Science Foundation

Award #ACI-1445604 Jetstream

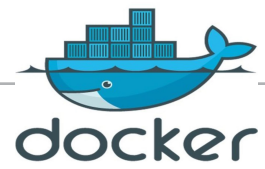
Award #ACI-1445606 Bridges

***XSEDE allocations: Pls may have support from
any funding agency or funding source.***

What is Jetstream?

- User-friendly, widely accessible cloud environment
- User-selectable library of preconfigured virtual machines; no need for system administration skills.
- NSF's first production cloud facility supporting all areas of science and engineering within NSF's scope
- Enable discoveries across disciplines such as biology, atmospheric science, economics, network science, observational astronomy, and social sciences.

Containers: Docker and Shifter



Shifter is built on docker, uses the images, etc. It replaces the run management (our end) to work with clusters and Cray. The repo retrieval is only slightly different (goes through a filter first), but can pull images made in docker and from dockerhub.

Basically, shifter is like docker bubble wrap for HPC systems, and people familiar with docker would have little to learn to run shifter: just little nuances.

Sheri Sanders

Shifter is "just another docker engine"... meaning it will run docker images... much the same way as the "original docker engine" runs docker images. Users can thus use all the existing docker containers from docker hub. The big picture: there is no issue using Shifter, even if all your work is in "the docker world".

Robert Henschel



Summarizing:

1) *The NCGAS Model*

1) *How to work with us:*

- i. We can host a server.*
- ii. We can provide a user interface, such as Galaxy*
- iii. We can “serve a community” as we do now for NSF researchers.*
- iv. We’re good at genomics/bioinformatics; image analyses would probably involve an additional RT group.*

Thank You

Questions?

Tom Doak (tdoak@iu.edu)

help@ncgas.org



**NATIONAL CENTER FOR
GENOME ANALYSIS SUPPORT**

INDIANA UNIVERSITY



INDIANA UNIVERSITY