

# Introduction to Research Data Management Using Globus



SDSC Summer Institute 2017

Rick Wagner

[rick@globus.org](mailto:rick@globus.org)

La Jolla, CA — August 4, 2017





# Research data management today



Index?

How do we...  
...move?  
...share?  
...discover?  
...reproduce?





Globus delivers...

Data transfer, sharing,  
publication, and discovery...

...directly from your own  
storage systems...

...via software-as-a-service



Globus enables...

# Campus Bridging

...within and beyond campus  
boundaries



# Bridge to campus HPC

Move datasets to campus research computing center



Move results to laptop, department, lab, ...



# Bridge to national cyberinfrastructure

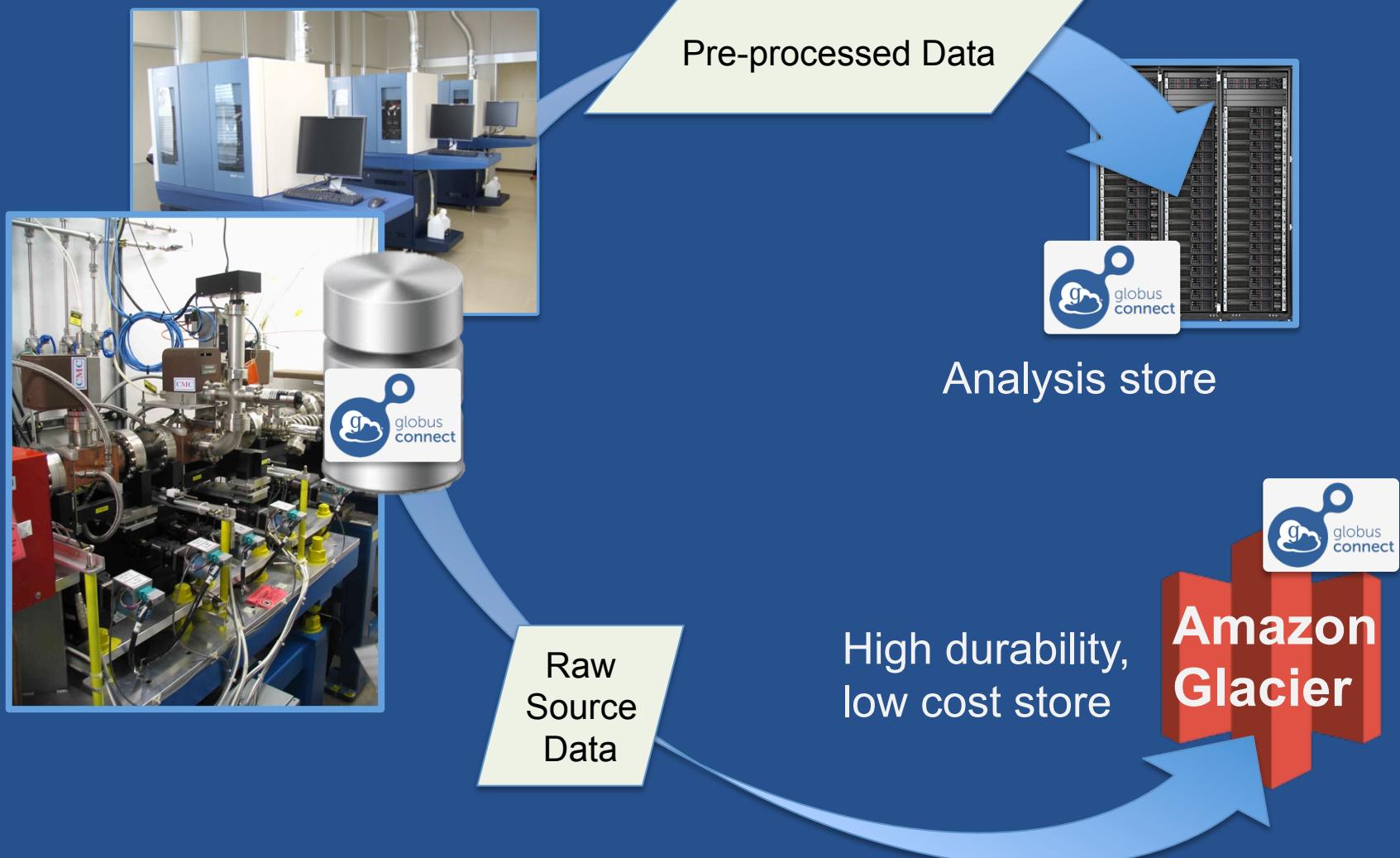
Move datasets to supercomputer,  
national facility



Move results to campus (...)

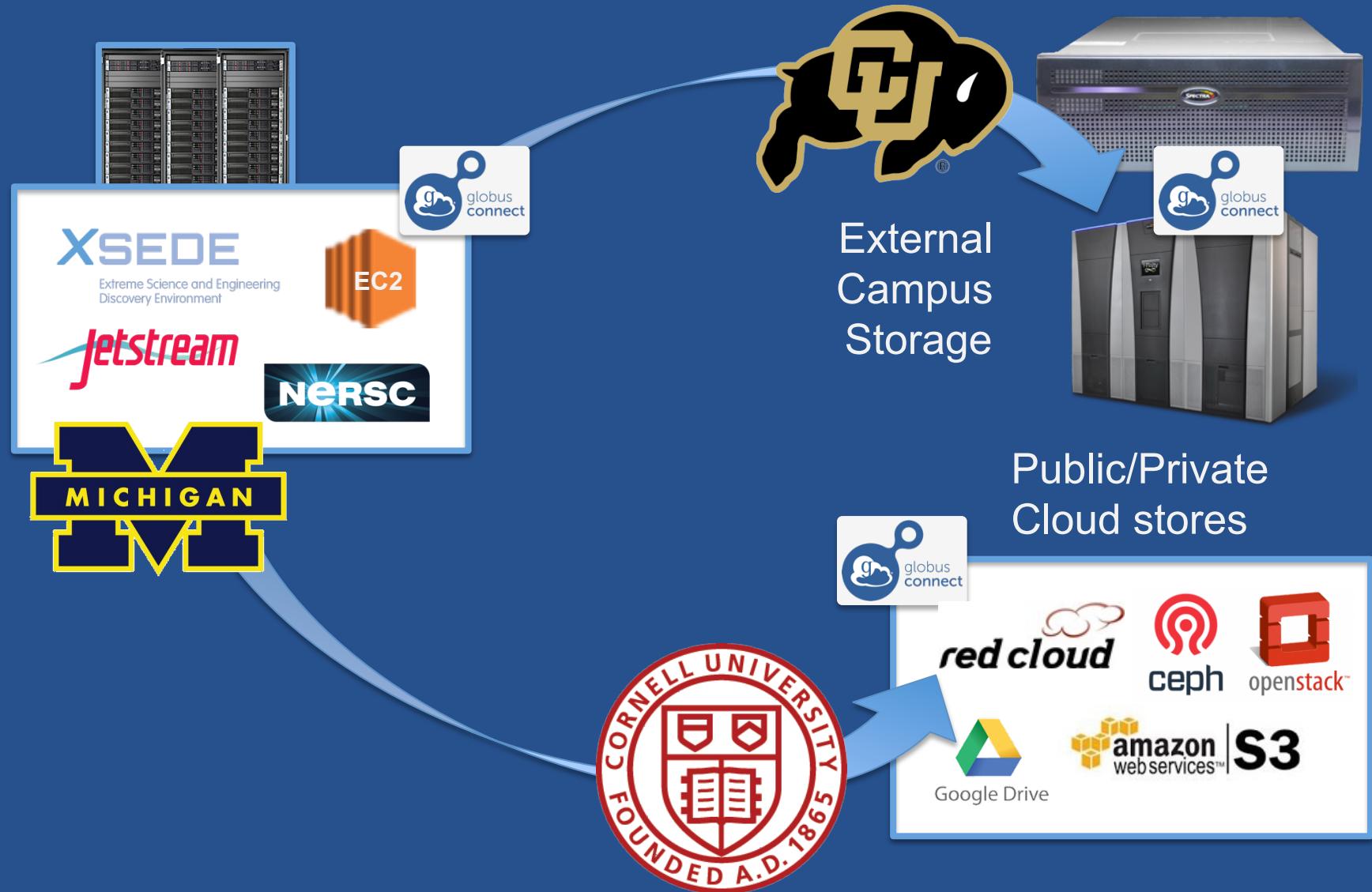


# Bridge to instruments





# Bridge to collaborators





# Bridge to community/public



Project Repositories,  
Replication Stores



Public Repositories





# Globus and the research data lifecycle

## Instrument



Globus transfers files  
reliably, securely

2

## Transfer

## Compute Facility



Researcher initiates transfer request; or requested automatically by script, science gateway

1



3

## Share

Researcher selects files to share, selects user or group, and sets access permissions

Collaborator logs in to Globus and accesses shared files; no local account required; download via Globus

5



## Personal Computer

4 Globus controls access to shared files on existing storage; no need to move files to cloud storage!

6

Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)

6

## Publish



Curator reviews and approves; data set published on campus or other system

7



## Publication Repository



## Discover

Peers, collaborators search and discover datasets; transfer and share using Globus

- Access via Web browser
- Use any storage system
- Use existing identity



# Why use Globus?

- **Simplicity**
  - Consistent UI across systems
  - Easy access to collaborators
- **Reliability and performance**
  - “Fire-and-forget” file transfer
  - Maximized WAN throughput
- **Operational efficiency**
  - Low overhead SaaS model
  - Highly automatable: CLI, RESTful API
- **Access to a large and growing community**



# Demonstration

# File Transfer



# How can I use Globus on my system?

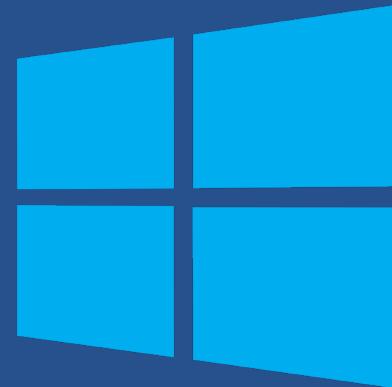
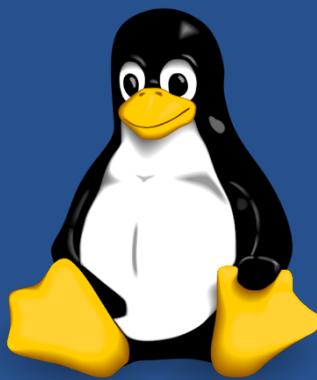


# Globus Connect...

Makes your storage  
system a Globus  
endpoint



# Globus Connect Personal



- **Installers do not require admin access**
- **Zero configuration; auto updating**
- **Handles NATs**



# Move files to/from your laptop

1. Go to: [www.globus.org/login](http://www.globus.org/login)
2. Select your institution from the list and click “Continue”
3. Authenticate with your institution’s identity system
4. Install Globus Connect Personal
5. Move file(s) between your laptop and another endpoint (e.g. ESnet)



# Demonstration

## File Sharing

## Federated Identity



# How can I integrate Globus into my research workflows?



# Globus serves as...

A platform for building science gateways, portals and other web applications in support of research and education



# Use(r)-appropriate interfaces



Globus service

The screenshot shows a file transfer interface. On the left, the 'xsede@longhorn' endpoint contains files like '000714\_Grid\_Healthcare\_KDM.rpt' (5.55 MB) and 'ClinicalHealth\_SummaryVP\_V1.docx' (13.64 kB). On the right, the 'esnet#ani-diskpt1' endpoint contains folders '10GB-in-small-files' and '50GB-in-small-files' with various subfiles. A central transfer progress bar indicates a transfer of '10GB.dat' from '10GB.dat' to '10GB.dat'.

Web

```
(globus-cli) jupiter:~ vas$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose           Control level of output
  -h, --help               Show this message and exit.
  -F, --format [json|text] Output format for stdout. Defaults to text
  --map-http-status TEXT  Map HTTP statuses to any of these exit codes: 0,
                           1, 50-99. e.g. "404=50,403=51" for the attack.
                           ...
  New Slide   Section   Help   Organizing the swarm.
  Section   Help   Bee 0 is joining the swarm.

Commands:
  bookmark      Manage Endpoint Bookmarks
  config        Modify, view, and manage your Globus CLI config.
```

CLI

```
GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
...
```

Rest  
API



# Globus as PaaS

XSEDE

Extreme Science and Engineering  
Discovery Environment



UNIVERSITY OF  
EXETER



Globus REST APIs

Data Publication & Discovery

File Sharing

File Transfer & Replication

Identity/Authentication,  
Group Management

Integrate file transfer and sharing capabilities into scientific web apps, portals, gateways, etc.

Globus Connect



XSEDE  
Extreme Science and Engineering  
Discovery Environment

SDSC

Enable existing institutional ID systems to be used in external web applications



# Data App: NCAR RDA

UCAR NCAR

Hello tuecke@uchicago.edu [dashboard](#) [sign out](#)

Closures/Emergencies Locations/Directions Find Pe

NCAR | UCAR Research Data Archive Computational & Information Systems Lab *weather • data • climate*

Go to Dataset: nnn.n

Home Find Data Ancillary Services About/Contact Data Citation Web Services For Staff

NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products  
ds094.2

For assistance, contact Bob Dattore (303-497-1825).

Description Data Access

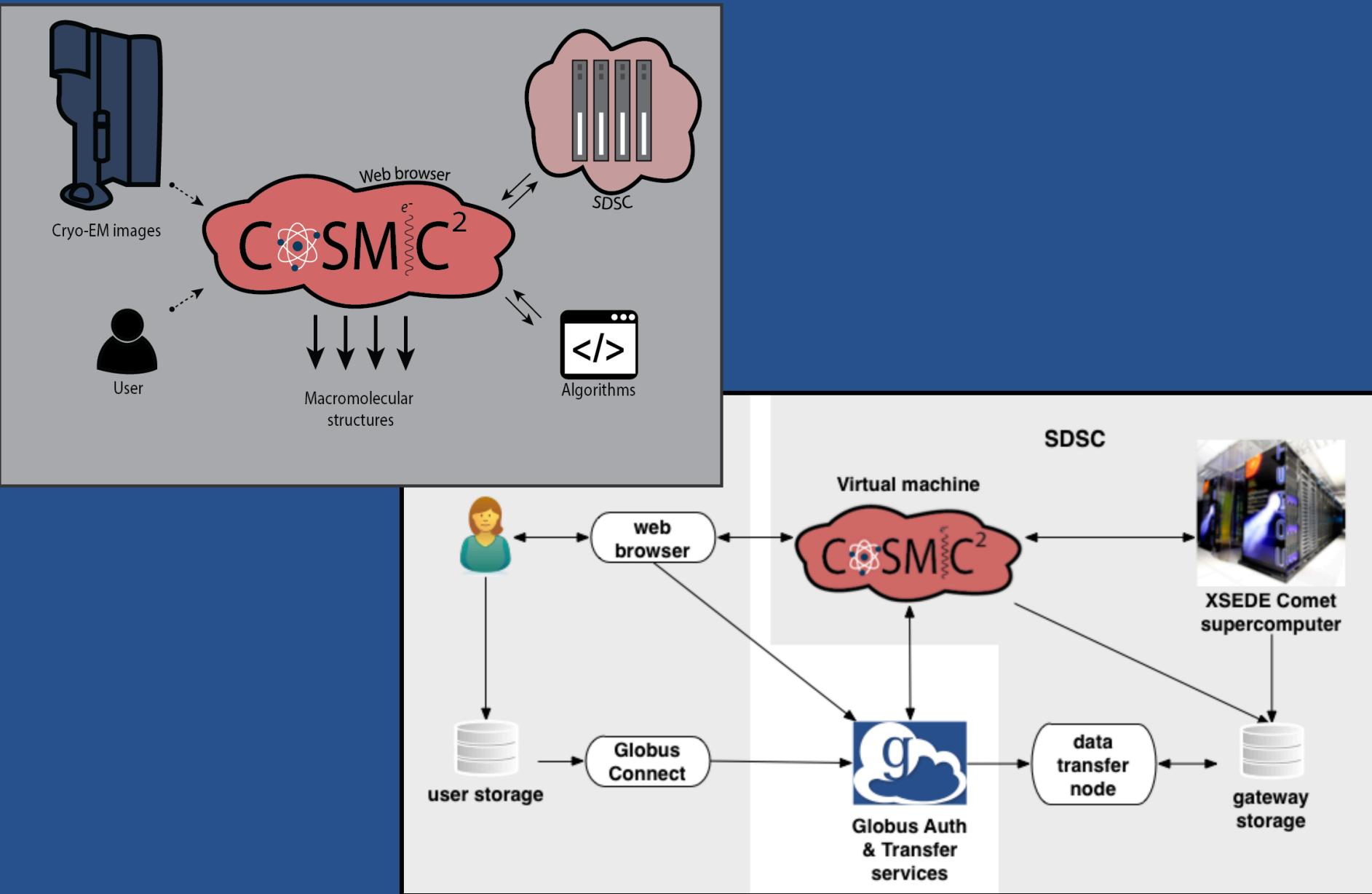
Mouse over the table headings for detailed descriptions

Data Description		Data File Downloads		Customizable Data Requests	Other Access Methods
		Web Server Holdings	Globus Transfer Service (GridFTP)	Subsetting	THREDDS Data Server
Union of Available Products		Web File Listing	Request Globus Invitation	Get a Subset	TDS Access
P	Diurnal monthly means	Web File Listing		Get a Subset	
R	Regular monthly means	Web File Listing		Get a Subset	

NCAR-Only Access	
Central File System (GLADE) Holdings	Tape Archive (HPSS) Holdings
GLADE File Listing	HPSS File Listing
GLADE File Listing	HPSS File Listing
GLADE File Listing	HPSS File Listing



# Gateway App: COSMIC<sup>2</sup> Cryo-EM





# Globus PaaS: National Resource Access

XSEDE  
Extreme Science and Engineering  
Discovery Environment

globus Account ▾

Jetstream Web App would like to:

Access all Jetstream resources

By clicking "Allow", you allow **Jetstream Web** information and services. You can rescind this

**Allow** **Deny**



compute canada | calcul canada

Globus Account Log In

Compute Canada has partnered with Globus to offer this high performance file transfer service.

Calcul Canada s'est associé à Globus pour vous offrir ce service de transfert de fichier à haute performance.

Log in to use Compute Canada Globus Web App

Use your existing organizational login  
e.g. university, national lab, facility, project, Google or [Globus ID](#)  
(Your Globus username and password used prior to February 13, 2016 is now Globus ID)

WestGrid

Continue

Didn't find your organization? Then use Globus ID to [sign up](#).



# Globus PaaS: Identity Management

The new Systems Biology Knowledgebase (KBase) is a collaborative effort designed to accelerate our understanding of microbes, microbial communities, and plants. It will be a community-driven, extensible and scalable open-source software framework and application system. KBase will offer free and open access to data, models and simulations, enabling scientists and researchers to build new knowledge and share their findings.

[Collaborate with us](#) [Get Started](#) [Develop with us](#)

**What can KBase do?**

- ✓ Combine heterogenous data types
- ✓ Offer standardized access to bioinformatic and modeling analyses
- ✓ Use evidence-supported annotations of genome structure and genetic function
- ✓ Discover new associations and network structures in community and molecular networks
- ✓ Map genotype to complex organismal traits
- ✓ Design and refine experiments using models of metabolism, regulation and community function
- ✓ Enable sharing of data, hypotheses, and newly-generated knowledge

**Latest News**

KBase at International Plant and Animal Genome XXI  
Posted by salazar Jan 09, 2013

KBase Team at Argonne for November Build  
Posted by nlharris Nov 30, 2012

November Build at Argonne  
Posted by salazar Nov 23, 2012

[view news](#)

**Upcoming Events**

2013-01-12  
International Plant and Animal Genome XXI (PAG 2013)

2013-02-18  
BERAC Presentations

2013-02-24  
DOE/NIFA Plant Feedstocks Genomics for Bioenergy

2013-02-25  
Proposed: Genomic Science Contractors-Grantees Meeting



# Globus PaaS: JupyterHub

jupyterhub / oauthenticator

Watch 14 Star 83 Fork 91

Code Issues 2 Pull requests 3 Projects 0 Wiki Insights

## Add Support for Globus OAuth #83

Merged minrk merged 8 commits into jupyterhub:master from NickolausDS:master on May 25

Conversation 5 Commits 8 Files changed 5 +401 -0

NickolausDS commented on May 24

The following adds support for Globus Auth, allowing users to login with their GlobusID. Alternatively, this authenticator supports other organizations supported by Globus Auth. For instance if someone wanted to setup their Jupyterhub server and restrict logins to the University of Chicago, they can configure their app to do so. The full list of institutions supported can be found [here](#).

In addition to Authentication, Globus also provides a transfer service for transferring data across nodes. This Authenticator is setup to procure `access_tokens` for users when they login, allowing them to easily transfer data from their python notebook with little hassle.

I followed the discussion [here](#) on storing tokens in `auth_state`, and in some outside discussion decided `auth_state` was the best place to do it. If security is a concern, admins can easily turn off transfer `access_tokens` in their `jupyterhub_config.py` file. By default, the Authenticator excludes user Auth `access_tokens`.

briedel and others added some commits on Mar 16

- Adding Globus Auth as a possible ID provider for jupyterhub 9284f8d
- Adding first version of tests 068c1f7
- Adding globus sdk as requirement d54326d
- Making requested changes. changing to use client\_id and client\_secret... 1cbcba...
- ...
- Changing requirements for globus f503dac
- Adding hosts for testing framework... hopefully these are right c7af663
- Added Globus features to globus.py oauthenticator ... 920ab74

Reviewers  
No reviews

Assignees  
No one assigned

Labels  
None yet

Projects  
None yet

Milestone  
No milestone

Notifications

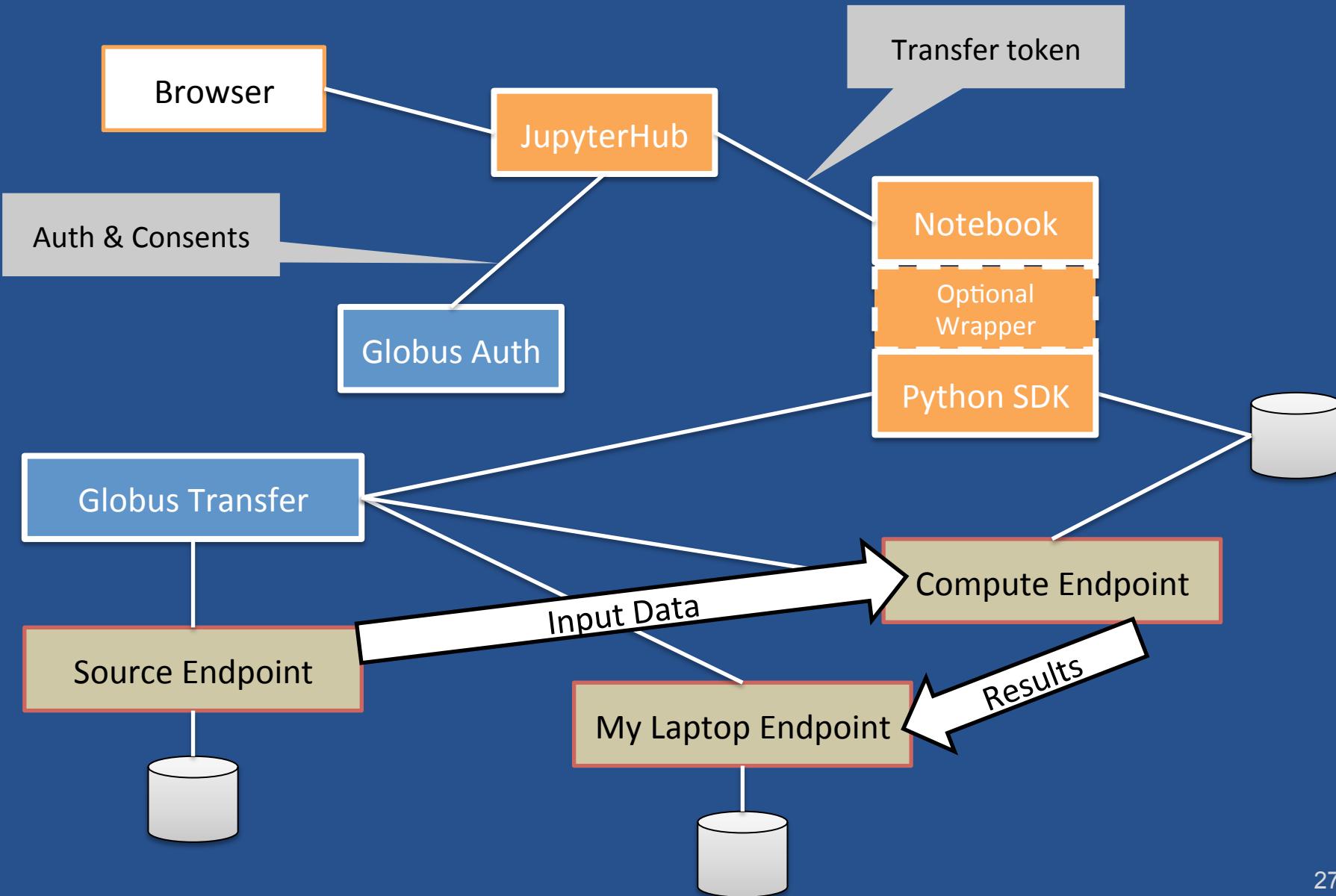
Subscribe

You're not receiving notifications from this thread.

3 participants



# Globus PaaS: JupyterHub



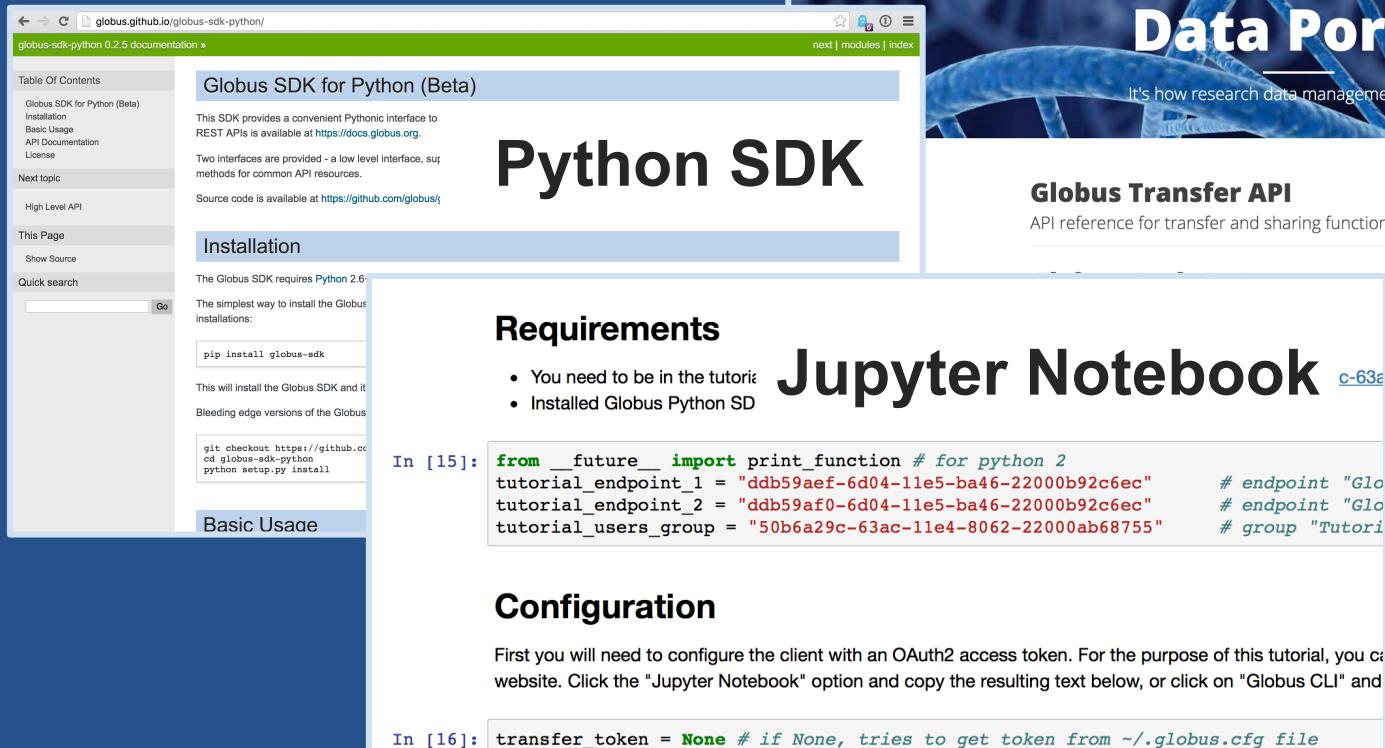


# Demonstration

# JupyterHub



# Globus PaaS developer resources



**Modern Research Data Portal**  
It's how research data management is done!

## Python SDK

### Requirements

- You need to be in the tutorial group
- Installed Globus Python SDK

```
In [15]: from __future__ import print_function # for python 2
tutorial_endpoint_1 = "ddb59aef-6d04-11e5-ba46-22000b92c6ec"           # endpoint "Global"
tutorial_endpoint_2 = "ddb59af0-6d04-11e5-ba46-22000b92c6ec"           # endpoint "Global"
tutorial_users_group = "50b6a29c-63ac-11e4-8062-22000ab68755"          # group "Tutorial"
```

### Configuration

First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can copy the token from the "Tutorial" group in the Globus UI. Go to the "My Groups" section of the website. Click the "Jupyter Notebook" option and copy the resulting text below, or click on "Globus CLI" and follow the instructions.

```
In [16]: transfer_token = None # if None, tries to get token from ~/.globus.cfg file
```

## Sample Application

[docs.globus.org/api](https://docs.globus.org/api)

[github.com/globus](https://github.com/globus)



# Thank you to our sponsors...



U.S. DEPARTMENT OF  
**ENERGY**



THE UNIVERSITY OF  
**CHICAGO**



**NIST**  
National Institute of  
**Standards and Technology**  
U.S. Department of Commerce



**Argonne**  
NATIONAL LABORATORY

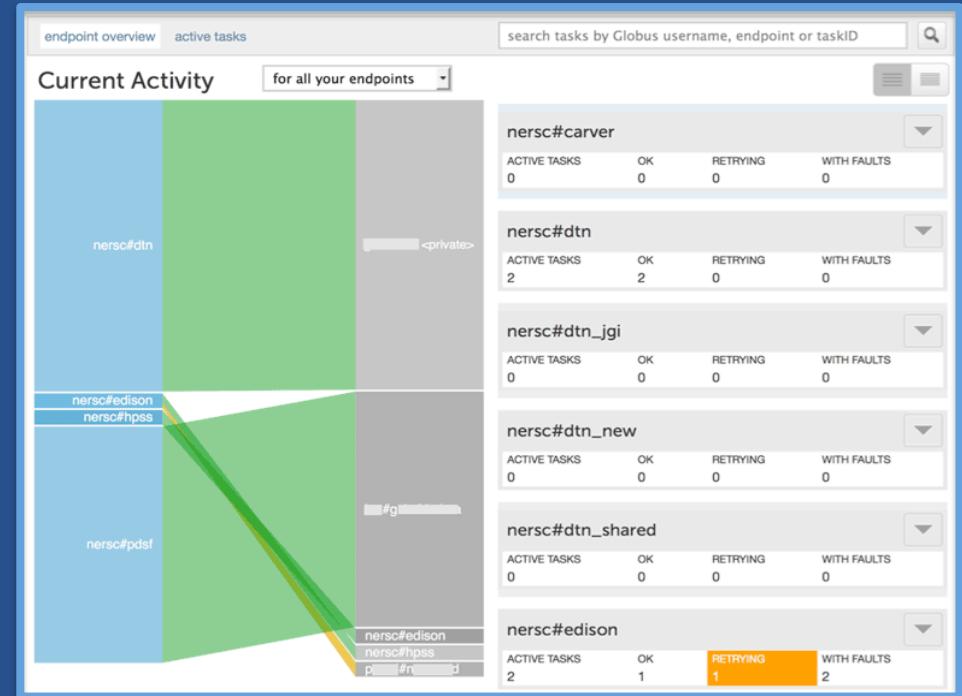
powered by  
**amazon**  
web services



# Globus sustainability model

- **Standard Subscription**

- Shared endpoints
- Data publication
- HTTPS support\*
- Management console
- Usage reporting
- Priority support
- Application integration



- **Branded Web Site**

- **Premium Storage Connectors**

- Amazon S3, Ceph, HPSS, Spectra, Google Drive, Box\*, HDFS\*

- **Alternate Identity Provider (InCommon is standard)**

\*Coming soon



# Thank you to our users...

**5**

major services

**290PB**

transferred

**47 Bn**

files processed

**70,000**

registered users

**13**

national labs  
use Globus

**10,000**

active endpoints

**10,000**

active users/year

**99.5%**

uptime

**65+**

institutional  
subscribers

**1 PB**

largest single  
transfer to date

**3 months**

longest  
continuously  
managed transfer

**300+**

federated  
campus identities



# ...and thank YOU, our subscribers!



NEW YORK UNIVERSITY



Berkeley  
UNIVERSITY OF CALIFORNIA



JOHNS HOPKINS  
UNIVERSITY

CORNELL  
UNIVERSITY



UF | UNIVERSITY of  
FLORIDA



MICHIGAN STATE  
UNIVERSITY

Yale



THE UNIVERSITY OF  
CHICAGO



SDSC

NIST

Virginia Tech  
*Invent the Future*





# Join the Globus community

- Access the service: [\*\*globus.org/login\*\*](https://globus.org/login)
- Create a personal endpoint: [\*\*globus.org/app/endpoints/create-gcp\*\*](https://globus.org/app/endpoints/create-gcp)
- Documentation: [\*\*docs.globus.org\*\*](https://docs.globus.org)
- Engage: [\*\*globus.org/mailing-lists\*\*](https://globus.org/mailing-lists)
- Subscribe: [\*\*globus.org/subscriptions\*\*](https://globus.org/subscriptions)
- Need help? [\*\*support@globus.org\*\*](mailto:support@globus.org)
- Follow us: [\*\*@globusonline\*\*](https://twitter.com/globusonline)