

# Managing Research Data with Globus Software-as-a-Service

ACI-REF Virtual Residency 2017

Vas Vasiliadis

[vas@uchicago.edu](mailto:vas@uchicago.edu)





# Research data management today



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

Index?





Globus delivers...

(Big) 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 research computing/HPC

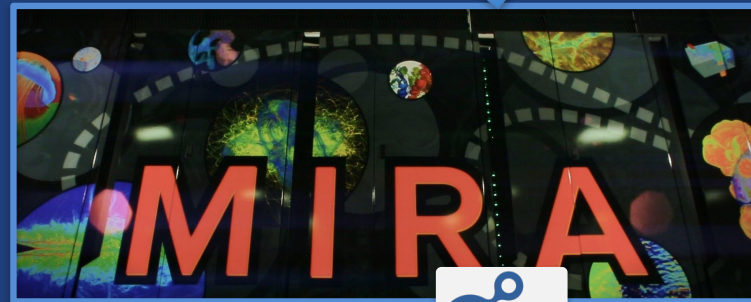
Move datasets to campus research computing center



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

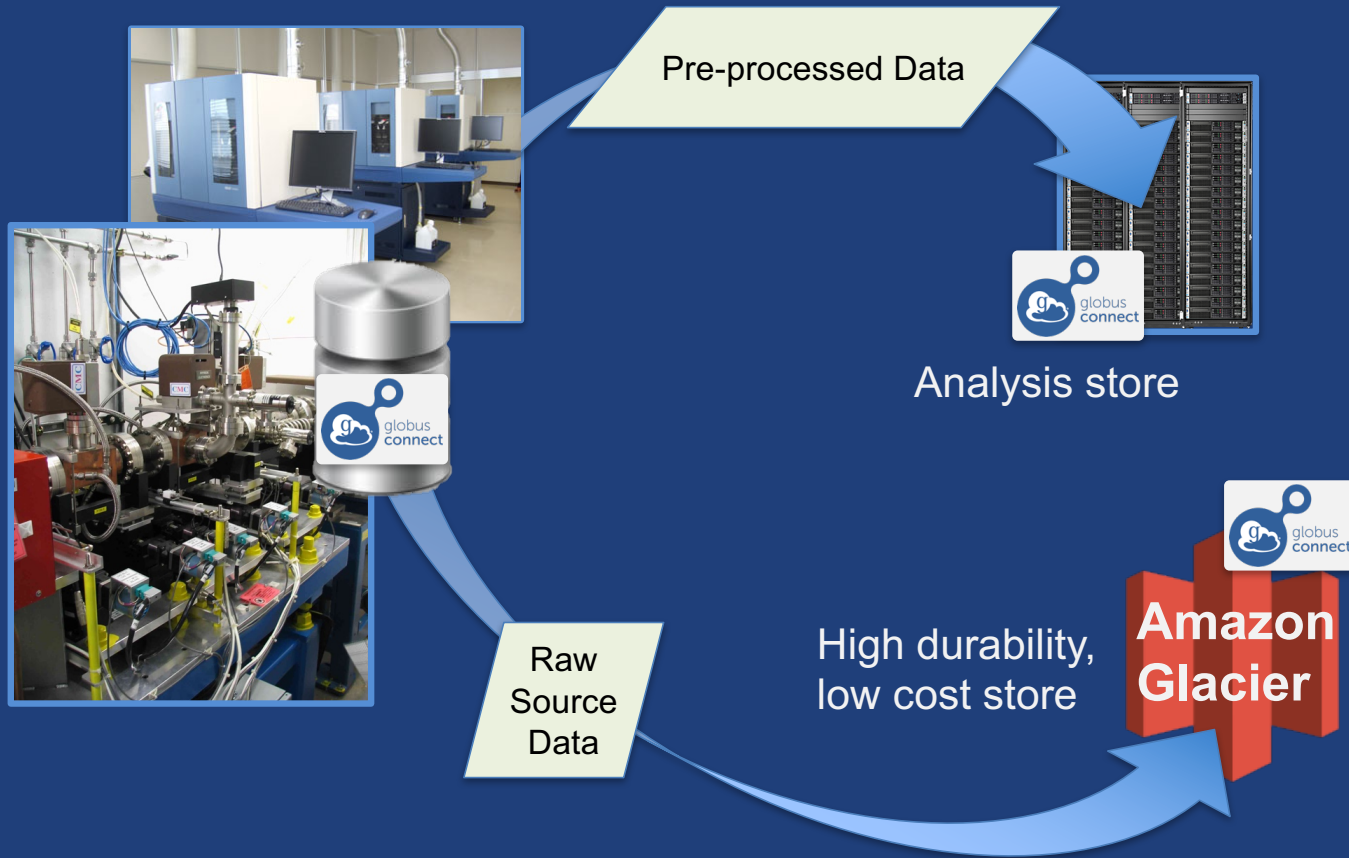
# Bridge to national cyberinfrastructure

Move datasets to supercomputer, national facility

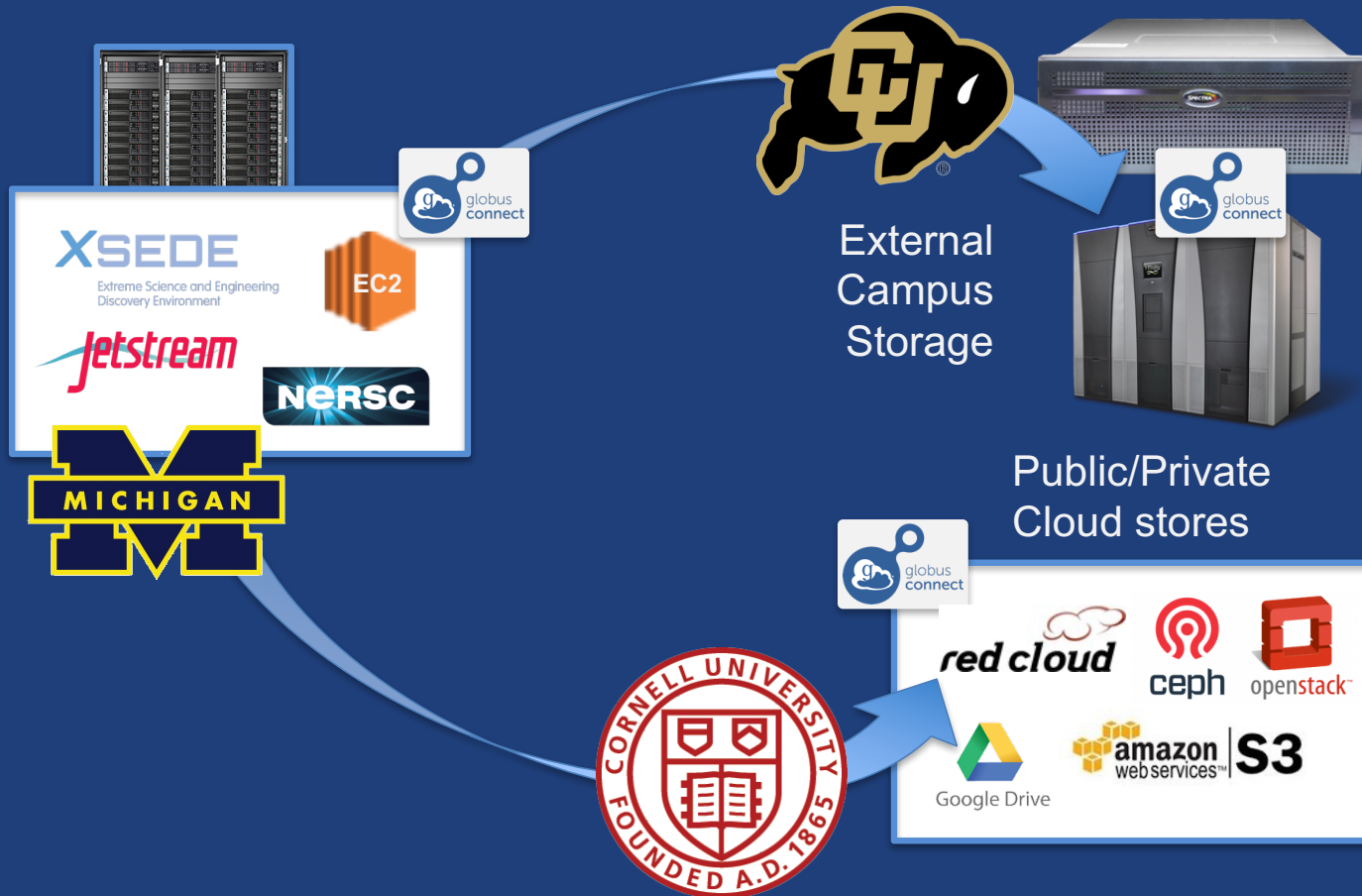


Move results to campus (...)

# Bridge to instruments



# Bridge to collaborators





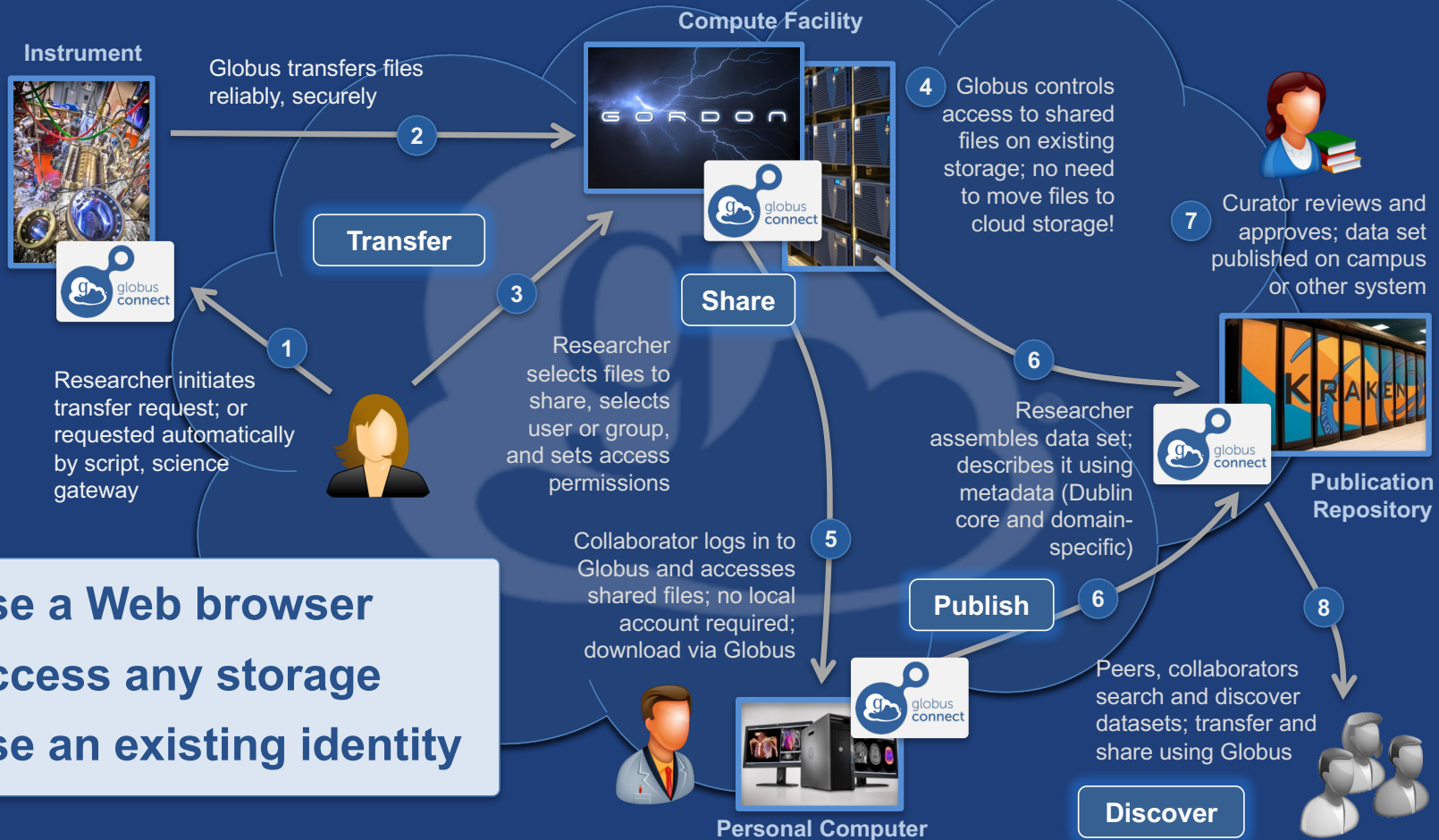


# Bridge to community/public





# Globus SaaS: Research data lifecycle



- Use a Web browser
- Access any storage
- Use an 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**



# Thank you to our users!

**48**

most server endpoints on one campus

**290 PB**  
transferred

**50 billion**  
tasks processed

**62,000**  
registered users

**350**

100TB+ users

**10,000**  
active users

**3 months**

longest running managed transfer

**10,000**

active endpoints

**350+**

federated identities

**1 PB**

largest single transfer to date

**5,119**

active shared endpoints

**99.5%**

uptime



# Demonstration

# File Transfer



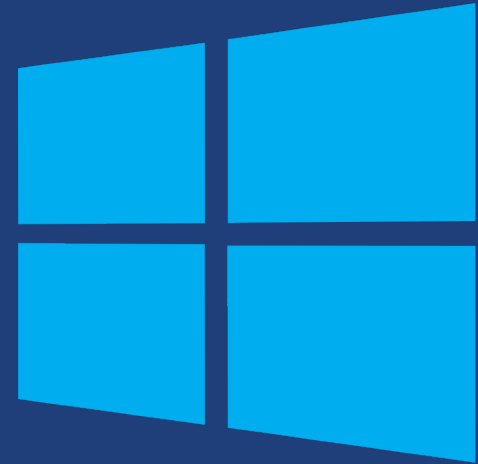
How can I use Globus  
on my computer?



**...makes your  
storage system a  
Globus endpoint**



# Globus Connect Personal



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





# Demonstration

## File Sharing

## Federated Identity



# Data Publication and Discovery

The screenshot shows the Materials Data Facility (MDF) community home page. At the top left is the Globus logo and the text "globus". To the right are "Log In" and "Sign Up" links. A blue banner below the header reads: "To submit a dataset or view datasets that have restricted access, please log in." Below this is a search bar with the placeholder text "Search" and a magnifying glass icon. The main content area features the title "Materials Data Facility Community home page" and a large logo consisting of a cluster of colorful circles (blue, green, yellow, orange, red) with the text "MATERIALS DATA FACILITY" overlaid. Below the logo, there is a paragraph of text: "The Materials Data Facility (MDF) is a scalable repository where materials scientists can publish, preserve, and share research data. The repository provides a focal point for the materials community, enabling publication and discovery of materials data of all sizes. MDF is a pilot project funded by NIST, and serves as the first pilot community of the [National Data Service](#). Contact Ben Blaiszik ([blaiszik@uchicago.edu](mailto:blaiszik@uchicago.edu)) to begin publishing your data". At the bottom, there is a "Browse" section with four buttons: "Issue Date", "Author", "Title", and "Subject".

<https://publish.globus.org>



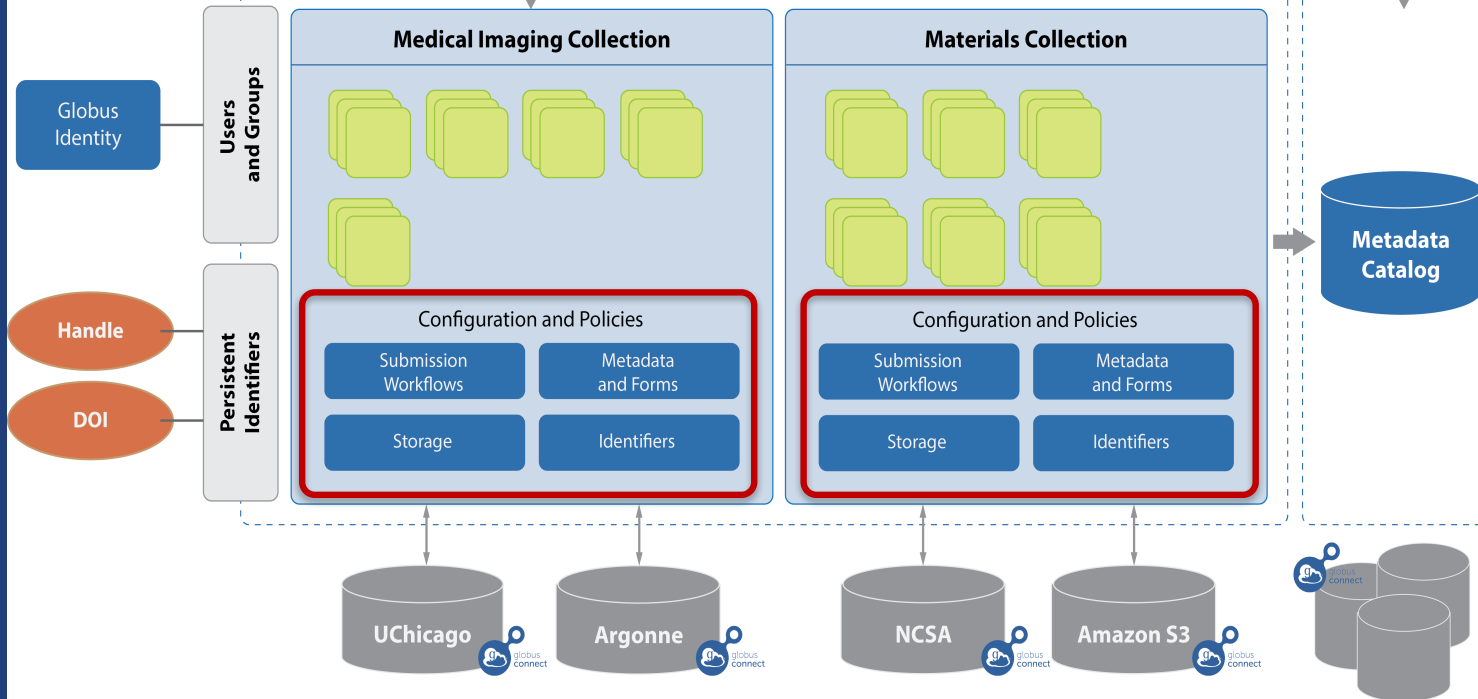
### Publish



### Discover

### Globus Authentication

### Globus Data Publication





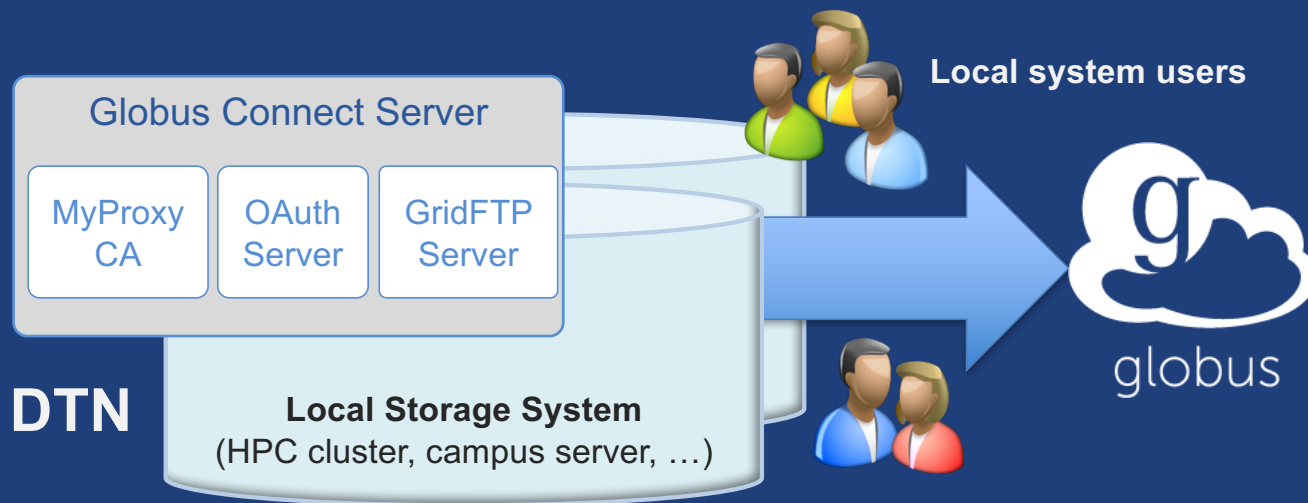
# Demonstration Data Publication



How do I do that  
on my research  
storage system(s)?



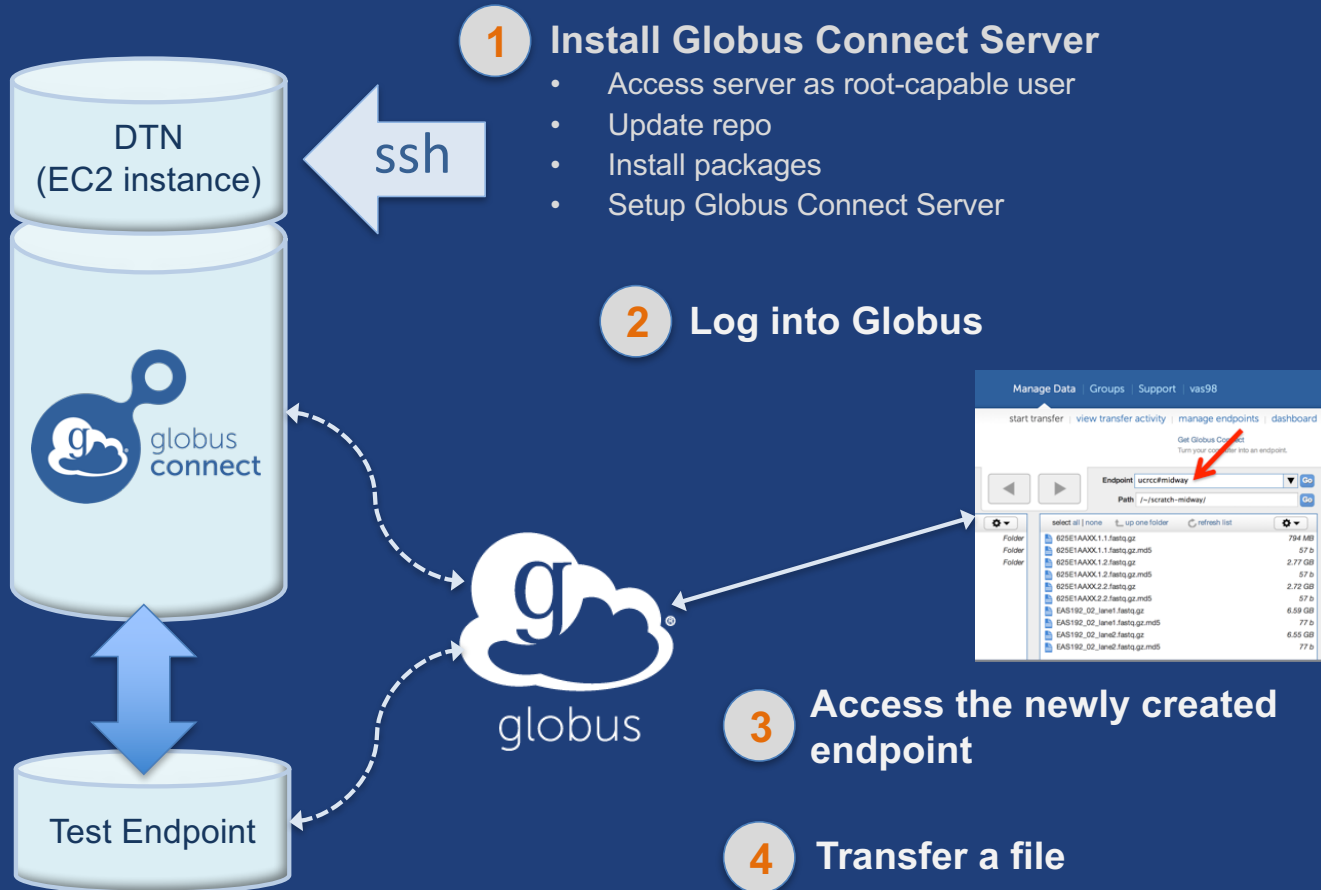
# Globus Connect Server



- **Create endpoint on practically any filesystem**
- **Enable access for all users with local accounts**
- **Native packages: RPMs and DEBs**



# Demonstration





## Installing Globus Connect Server

```
$ sudo su
$ curl -LOs http://toolkit.globus.org/ftppub/globus-
connect-server/globus-connect-server-
repo_latest_all.deb
$ dpkg -i globus-connect-server-repo_latest_all.deb
$ apt-get update
$ apt-get -y install globus-connect-server
$ globus-connect-server-setup
$ _
```

**You have a working Globus endpoint!**





## Common configuration options

- **Manage Endpoints page**
  - Display Name
  - Visibility
  - Encryption
- **DTN local config file (.ini format)**
  - `/etc/globus-connect-server.conf`
  - RestrictPaths
  - IdentityMethod (CILogon, OAuth)
  - Sharing
  - SharingRestrictPaths



# Storage connectors

- **Standard storage connectors (POSIX)**

- Linux, Windows, MacOS
- Lustre, GPFS, OrangeFS, etc.

- **Premium storage connectors**

AWS S3

Ceph RadosGW (S3 API)

Spectra Logic BlackPearl

HPSS

Google Drive (beta)

Box (in progress)

HDFS (in progress)

iRODS (in progress)

HGST Active Archive (in progress)

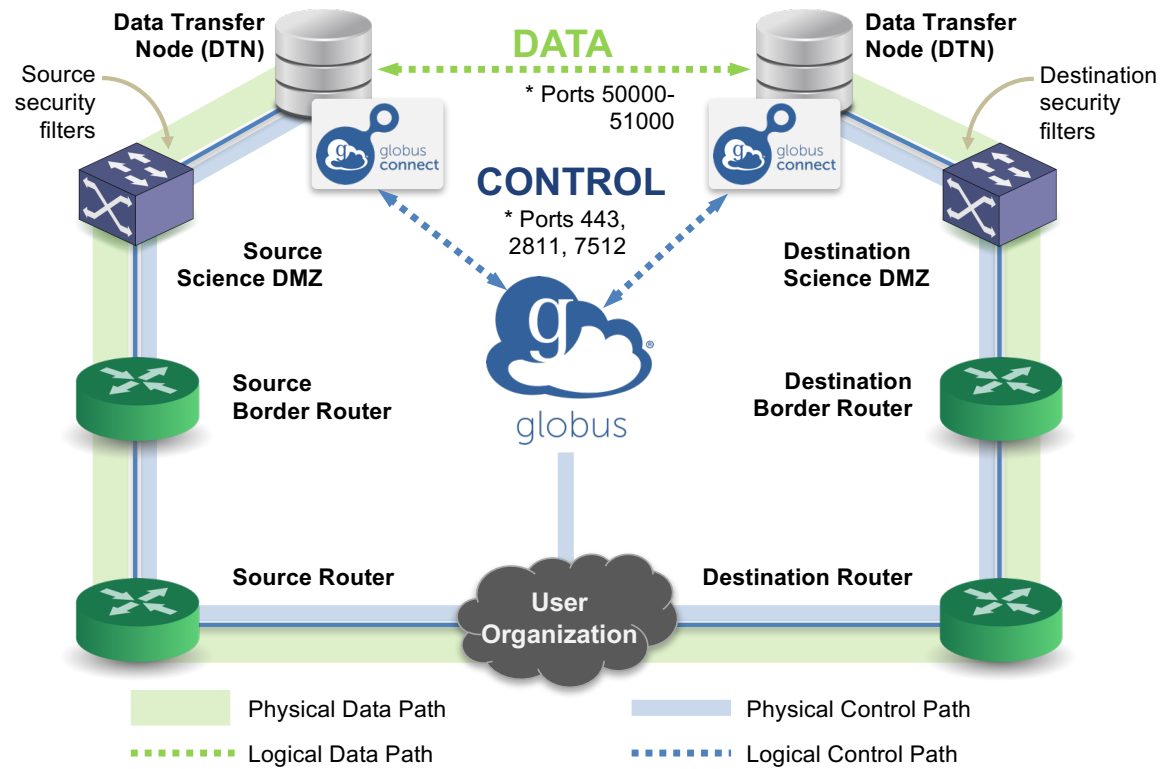
[docs.globus.org/premium-storage-connectors](https://docs.globus.org/premium-storage-connectors)



# Deployment Best Practices



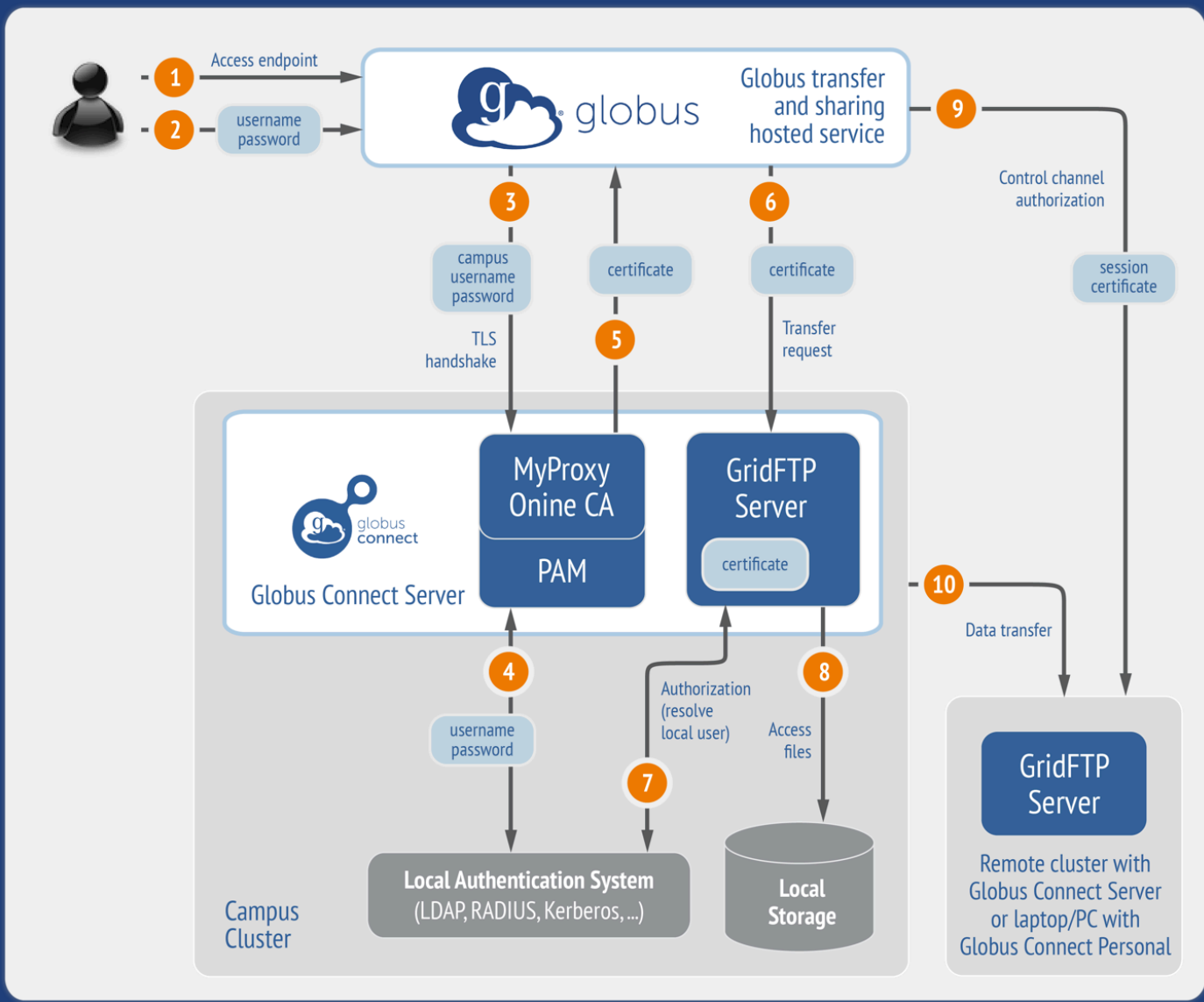
# Data Transfer Node in a Science DMZ



\* Please see TCP ports reference: [https://docs.globus.org/resource-provider-guide/#open-tcp-ports\\_section](https://docs.globus.org/resource-provider-guide/#open-tcp-ports_section)



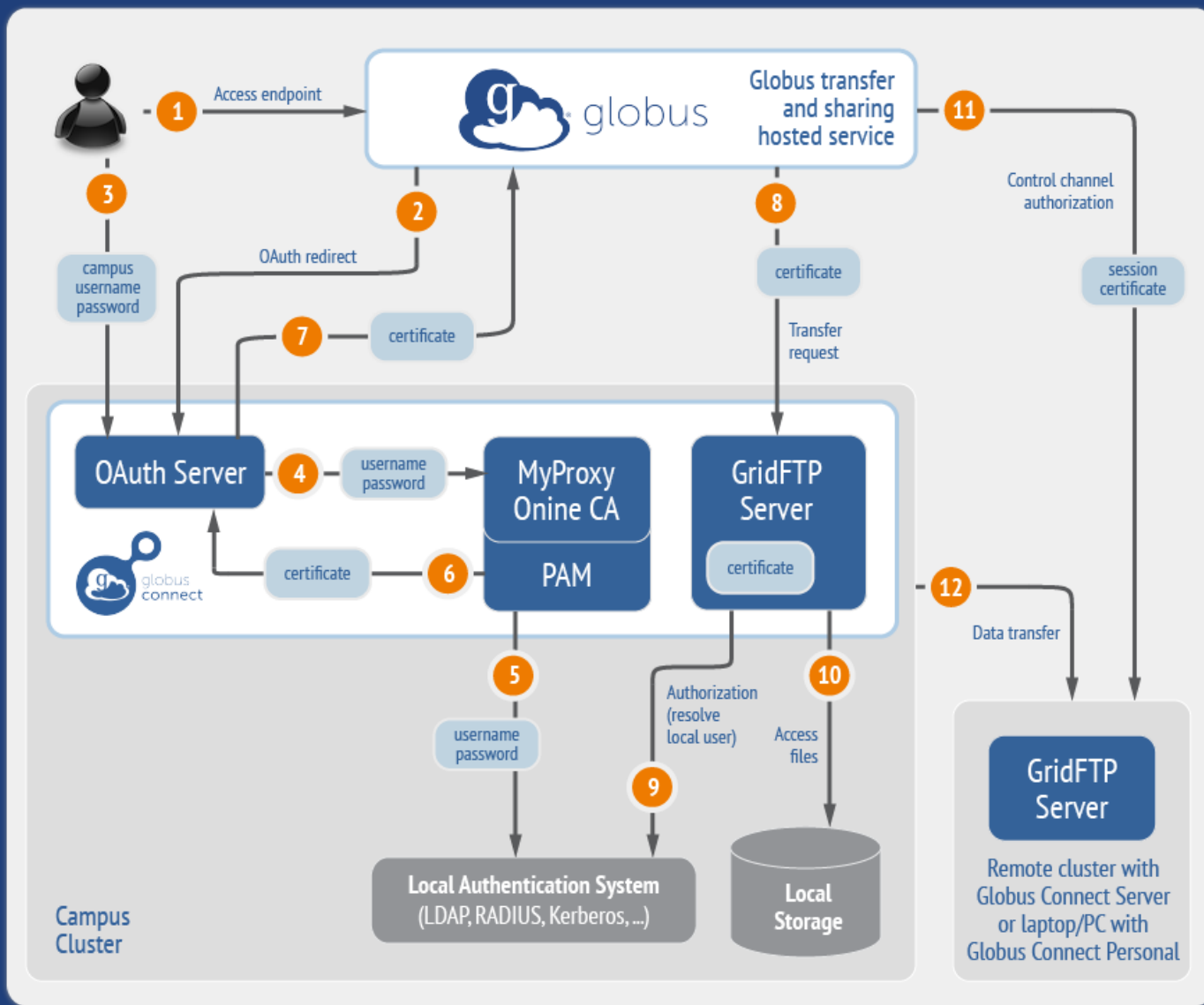
# Endpoint activation using MyProxy



**DON'T  
LEAVE IT  
LIKE THIS!**



# Endpoint activation using MyProxy OAuth



Yes, please do this!



**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**

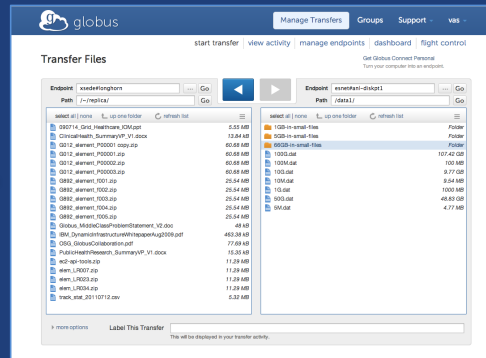




# Use(r)-appropriate interfaces



Globus service



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"

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 Command Line Interface

- Transfer and Auth
- Uses Python SDK
- Open source

[github.com/globus/globus-cli](https://github.com/globus/globus-cli)  
[docs.globus.org/cli](https://docs.globus.org/cli)

```
(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"

Commands:
  bookmark      Manage Endpoint Bookmarks
  config        Modify, view, and manage your Globus CLI config.
  delete        Submit a Delete Task
  endpoint      Manage Globus Endpoint definitions
  get-identities Lookup Globus Auth Identities
  list-commands List all CLI Commands
  login         Login to Globus to get credentials for the Globus CLI
  logout        Logout of the Globus CLI
  ls            List Endpoint directory contents
  mkdir         Make a directory on an Endpoint
  rename        Rename a file or directory on an Endpoint
  task          Manage asynchronous Tasks
  transfer      Submit a Transfer Task
  version       Show the version and exit
  whoami        Show the currently logged-in identity.
```

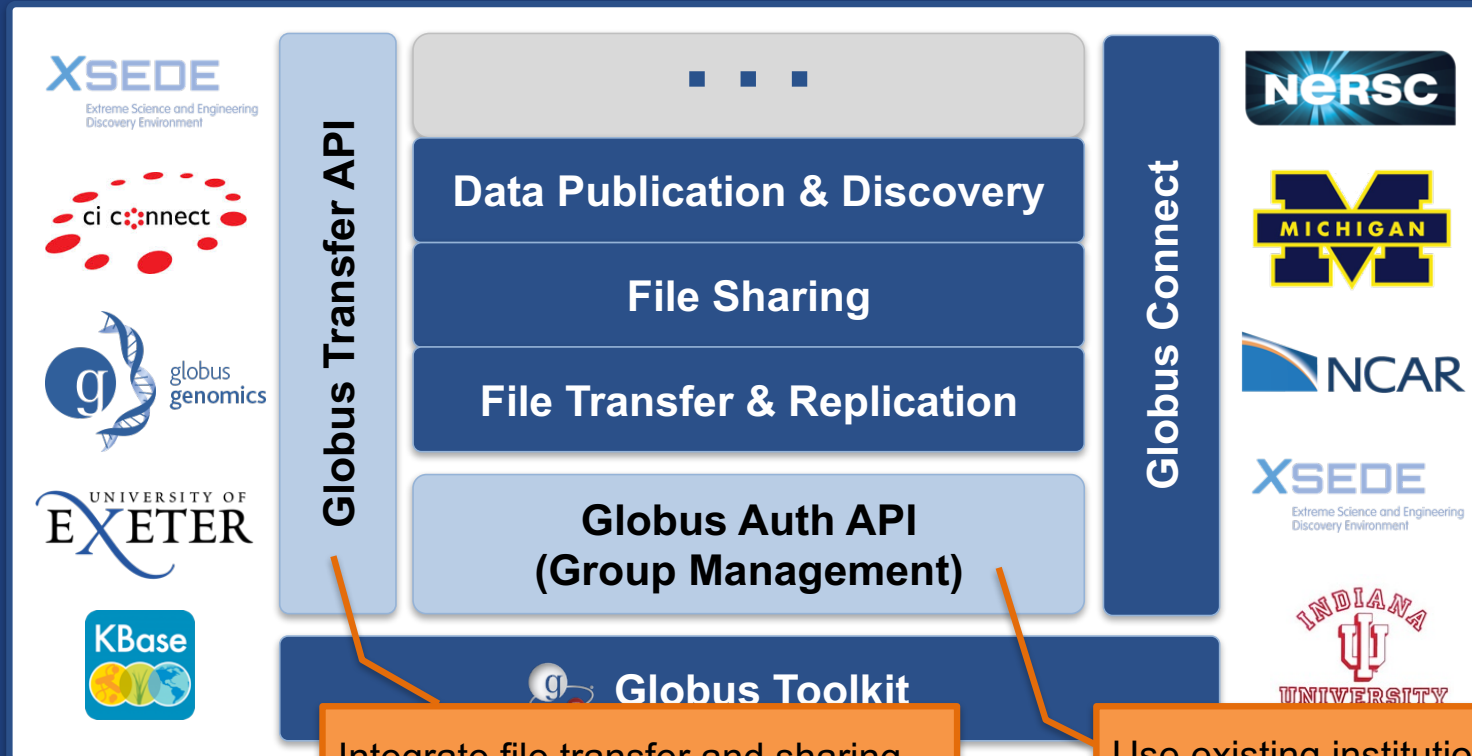


# Demonstration

# Command Line Interface



# Globus as PaaS



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

Use existing institutional ID systems in external web applications

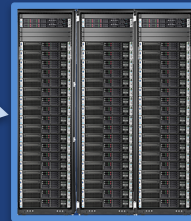


# Simple Automation

## Recurring transfers with sync option



Copy /ingest  
Daily @ 3:30am



## Data distribution

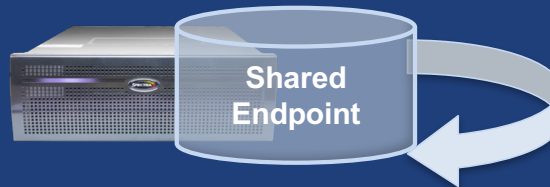


.../my\_share

- /cohort045
- /cohort096
- /cohort127



## Staging area cleanup




1. Check if successful transfer
2. Delete data from staging area



# Data Portal: NCAR RDA

UCAR NCAR Closures/Emergencies Locations/Directions Find Pe

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

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

Go to Dataset:

[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).

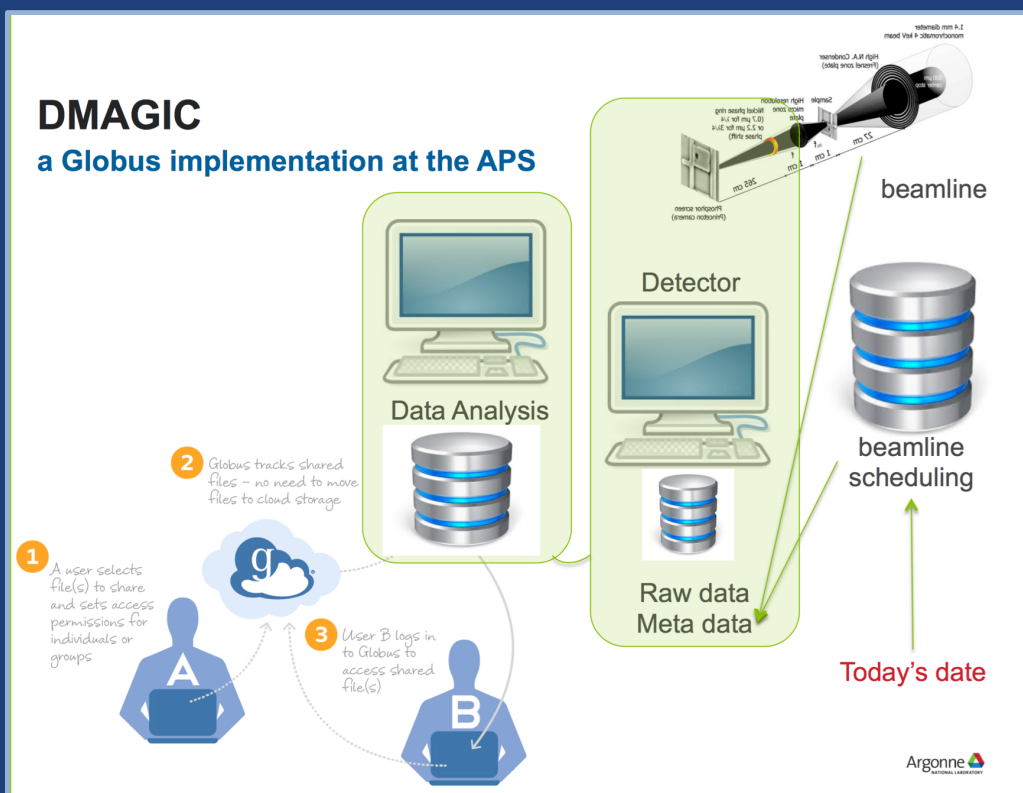
[Description](#) [Data Access](#)

Mouse over the table headings for detailed descriptions

Data Description		Data File Downloads		Customizable Data Requests	Other Access Methods	NCAR-Only Access	
		Web Server Holdings	Globus Transfer Service (GridFTP)	Subsetting	THREDDS Data Server	Central File System (GLADE) Holdings	Tape Archive (HPSS) Holdings
Union of Available Products		Web File Listing	Request Globus Invitation	Get a Subset	TDS Access	GLADE File Listing	HPSS File Listing
P R O D	Diurnal monthly means	Web File Listing		Get a Subset		GLADE File Listing	HPSS File Listing
	Regular monthly means	Web File Listing		Get a Subset		GLADE File Listing	HPSS File Listing



# Data Distribution; APS - DMagic



[dmagic.readthedocs.io](https://dmagic.readthedocs.io)

DMagic  
latest

Search docs

About DMagic  
Install directions  
Development  
API reference  
Examples  
Frequently asked questions

Docs » DMagic [Edit on GitHub](#)

DMagic is an open-sourced Python toolbox to perform data management and data sharing for users of the Imaging Group of the Advanced Photon Source.

Courtesy of Francesco De Carlo, Argonne National Laboratory (2016)



# Analysis App: Wellcome Sanger

Sanger Imputation Service **Beta**

Home About Instructions Resources Status

## Sanger Imputation Service

This is a free genotype **imputation** and **phasing** service provided by the [Wellcome Trust Sanger Institute](#). You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click [here](#) to learn more and [follow us on Twitter](#).

### Before you start

Be sure to [read through the instructions](#).  
You will need to set up a free account with [Globus](#) and have [Globus Connect](#) running at your institute or on your computer to transfer files to and from the service.

### Ready to start?

If you are ready to upload your data, please fill in the details below to **register an imputation and/or phasing job**. If you need more information, see the [about](#) page.

What is this

[Next](#)

### News

[@sangerimpute](#)

**11/05/2016**

Thanks to [EAGLE](#), we can now return **phased data**. The HRC panel has been updated to r1.1 to fix a [known issue](#). See [ChangeLog](#) for more details.

**15/02/2016**

Globus API changed, please see [updated instructions](#).

**17/12/2015**

New status page and reworked internals. See [ChangeLog](#).

**09/11/2015**

Pipeline updated to add some features requested by users. See [ChangeLog](#).

[See older news...](#)





# National Resource Access

**XSEDE**  
Extreme Science and Engineering  
Discovery Environment

globus Account ▾

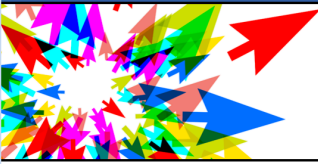
Jetstream Web App would like to:

✔ Access all Jetstream resources

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

**Allow** **Deny**

globus Globus Account Log In



**compute** | **calcul**  
canada | canada

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

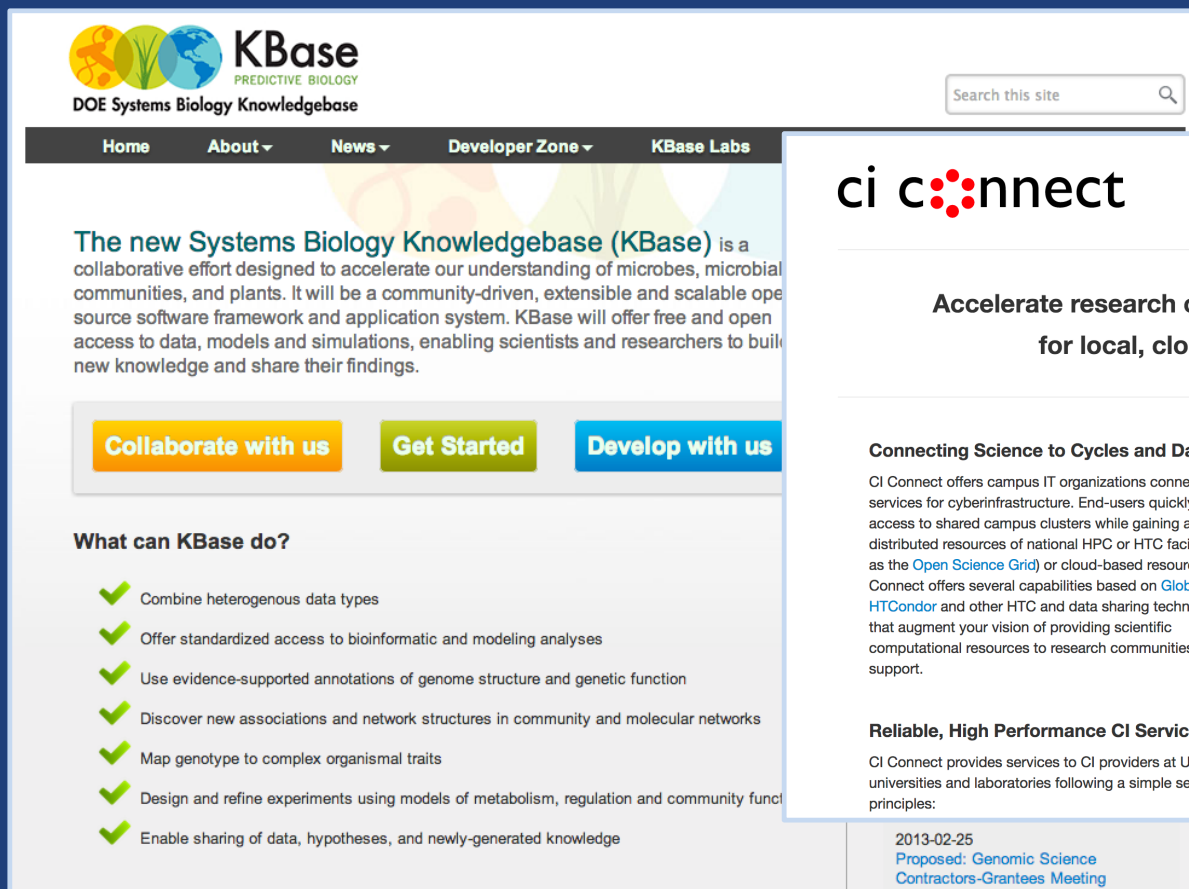
e.g. university, national lab, facility, project, Google or [Globus ID](#)  
(Your Globus username and password used prior to February 13, 2016 is now Globus ID)

WestGrid ▾

**Continue**

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

# Identity Management



The screenshot shows the KBase website homepage. At the top left is the KBase logo with the text "KBase PREDICTIVE BIOLOGY DOE Systems Biology Knowledgebase". To the right is a search bar labeled "Search this site". Below the logo is a navigation menu with "Home", "About", "News", "Developer Zone", and "KBase Labs". The main content area features a large heading "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." Below this are three buttons: "Collaborate with us", "Get Started", and "Develop with us". A section titled "What can KBase do?" lists seven bullet points with green checkmarks: "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", and "Enable sharing of data, hypotheses, and newly-generated knowledge".

## ci connect

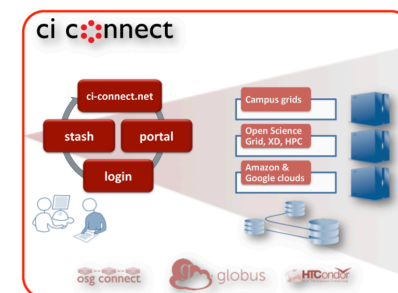
Accelerate research on campus by providing connective services for local, cloud and national cyberinfrastructure

### Connecting Science to Cycles and Data

CI Connect offers campus IT organizations connective services for cyberinfrastructure. End-users quickly gain access to shared campus clusters while gaining access to distributed resources of national HPC or HTC facilities (such as the [Open Science Grid](#)) or cloud-based resources. CI Connect offers several capabilities based on [Globus](#), [HTCondor](#) and other HTC and data sharing technologies that augment your vision of providing scientific computational resources to research communities you support.

### Reliable, High Performance CI Services

CI Connect provides services to CI providers at US universities and laboratories following a simple set of principles:



### Connected environments from hosted services

Resources of a campus cluster (or [campus grid](#)) can be

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



# Globus PaaS developer resources



globus.github.io/globus-sdk-python/

globus-sdk-python 0.2.5 documentation » next | modules | index

Table Of Contents

- Globus SDK for Python (Beta)
- Installation
- Basic Usage
- API Documentation
- License

## Python SDK

Next: globus/globus-sdk-python.

This Page

- Show Source
- Quick search

### Installation

The Globus SDK requires Python 2.6+ or 3.2+. If a successful installation is not possible, please refer to the installation instructions.

```
pip install globus-sdk
```

This will install the Globus SDK and its dependencies.

Bleeding edge versions of the Globus SDK can be installed using the following instructions:

```
git checkout https://github.com/globus/globus-sdk-python
cd globus-sdk-python
python setup.py install
```

### Basic Usage

## Requirements

- You need to be in the tutorial users group for sharing: <https://www.globus.org/app/groups/50b6a29c-63ac-11e4-8062-22000ab68755>
- Installed Globus Python SDK

## Jupyter Notebook

```
In [15]: from __future__ import print_function
tutorial_endpoint_1 = "ddb59ae1-0004-11e3-ba46-22000b92c6ec" # endpoint "Globus"
tutorial_endpoint_2 = "ddb59af0-6d04-11e5-ba46-22000b92c6ec" # endpoint "Globus"
tutorial_users_group = "50b6a29c-63ac-11e4-8062-22000ab68755" # group "Tutorial Users Group"
```

## Configuration

First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can use the Globus CLI to generate a token. Visit the Globus CLI website. Click the "Jupyter Notebook" option and copy the resulting text below, or click on "Globus CLI" and

```
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)



**...on sustainability**



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



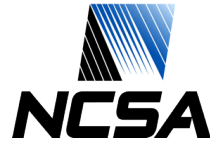
**Argonne**  
NATIONAL LABORATORY



powered by  
**amazon**  
web services



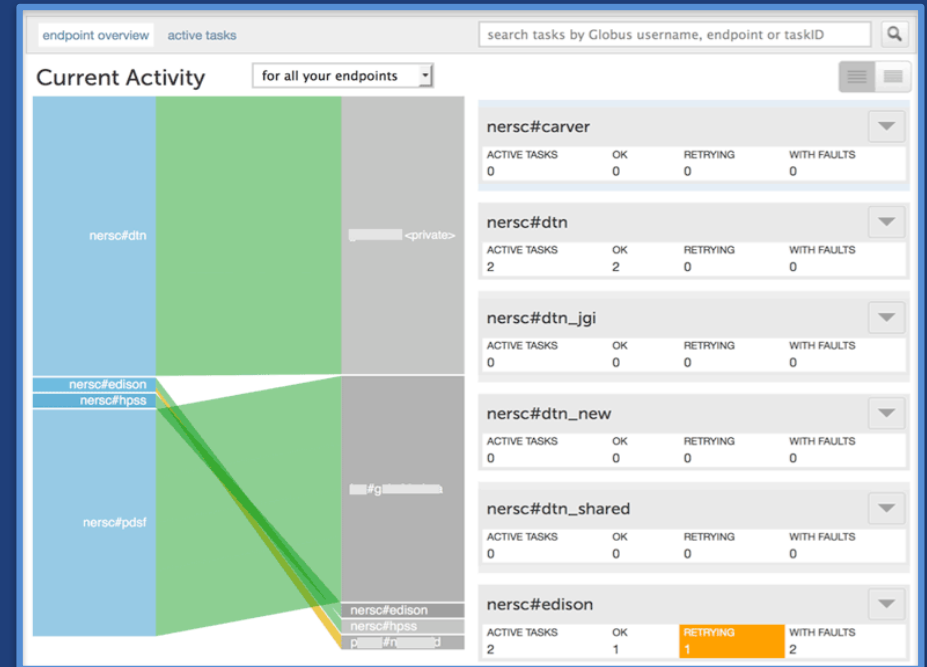
...and THANK YOU, subscribers!





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



\*Coming soon



# Demonstration Management Console



## Facilitating adoption

- **GlobusWorld Tour:**  
7 stops, over 250 attendees
- **Lots of open source  
developer materials**
- **Contact us to host event**



**September 12-13**



**Yale**



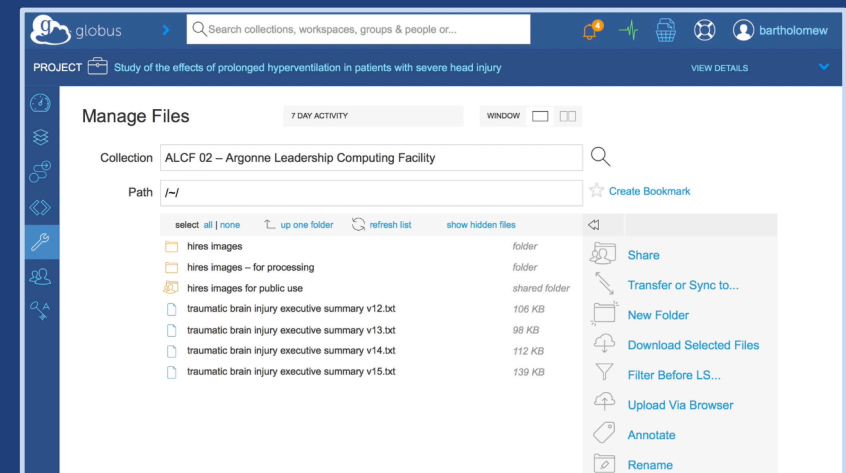
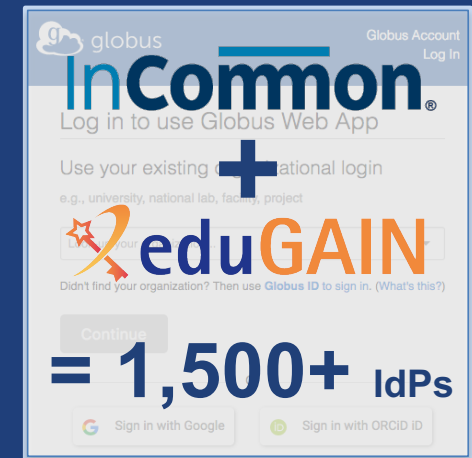
**PennState**



**NEW YORK UNIVERSITY**

# Future Directions

- New data management interface
- Collections: the evolution of endpoints
- HTTPS access to storage
- Enhanced sharing capabilities
- Metadata and search
- Data Publication PaaS





## 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)