

# Data Mobility Perspectives and the Modern Research Data Portal

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

Data Mobility Workshop  
September 23, 2019





## MRDP: Key elements

**Science DMZ**  
Fast, clean data path

**Data Transfer Node**  
Purpose-built data mover

**Globus Platform**  
Secure, reliable data  
orchestration

**Globus Connect**  
Storage system enabler



# Who/What/Where is Globus?

## Data mobility perspectives...

- Researcher
- Service provider/sysadmin
- Portal/app developer



...a brief detour 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!



JOHNS HOPKINS  
UNIVERSITY



NCAR

Yale



wellcome trust  
**sanger**  
institute



HARVARD  
UNIVERSITY



UNIVERSITY of  
**FLORIDA**

**CORNELL**  
UNIVERSITY



NEW YORK UNIVERSITY



THE UNIVERSITY OF  
**CHICAGO**



MICHIGAN STATE  
UNIVERSITY



VirginiaTech  
*Invent the Future*



Dartmouth

**syngenta**

**NLS**

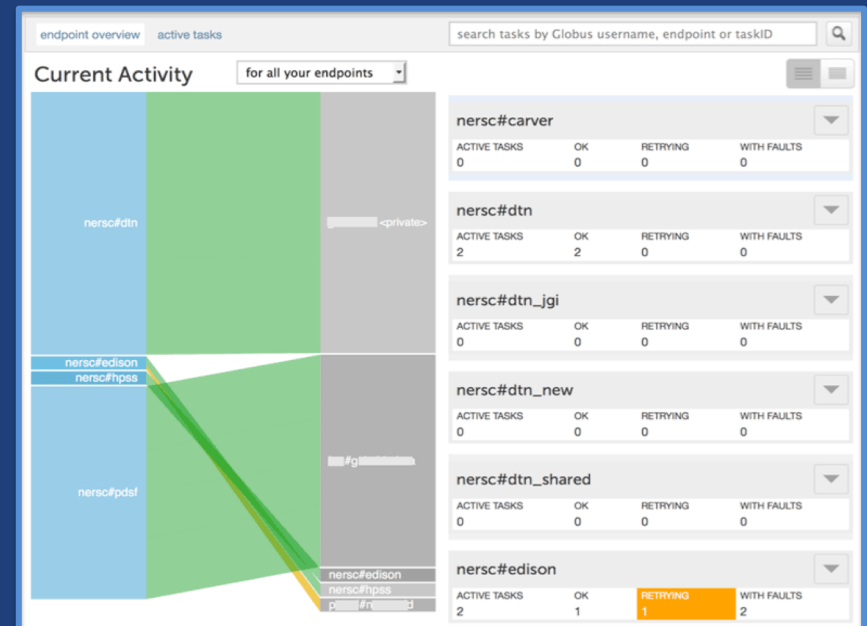
SIMONS FOUNDATION





# Globus sustainability model

- **Standard Subscription**
  - Sharing, data publication
  - HTTPS access
  - Console, usage reporting
  - Priority support
  - App integration support
- **High Assurance subscription**
  - App instance isolation
  - Additional authentication assurance
  - Audit logging
  - NIST 800-53, NIST 800-171 (+ BAA)
- **Branded Web Site**
- **Premium Storage Connectors**
- **Alternate Identity Provider (InCommon is standard)**



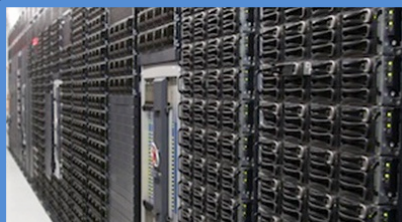


# The Researcher Perspective

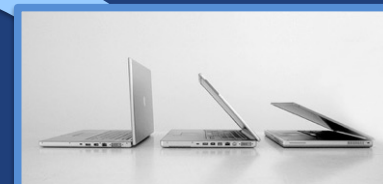




# Unified data access across diverse storage systems



Research Computing HPC



Personal Resources



Desktop Workstations



Mass Storage

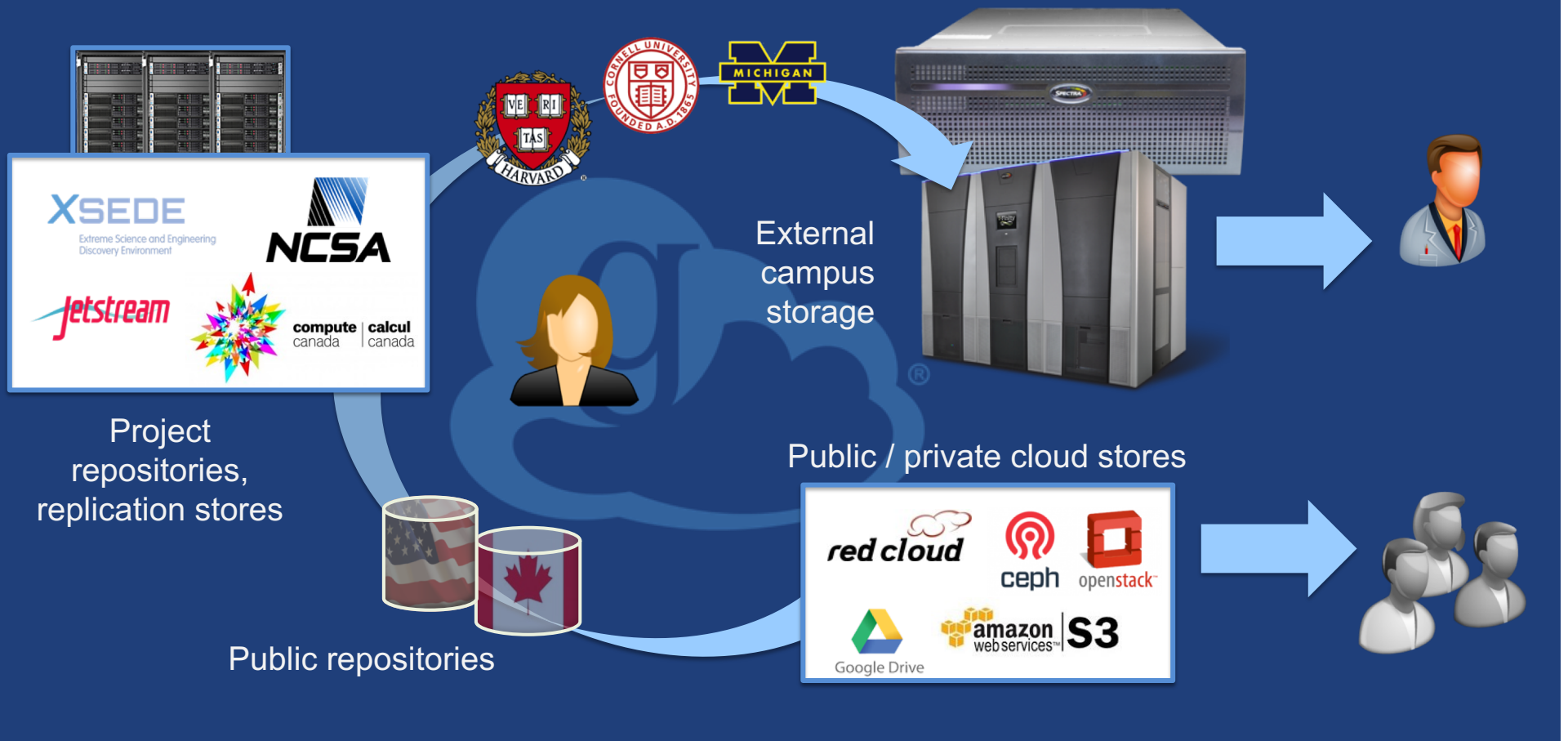


Instruments





# Sharing with collaborators, community

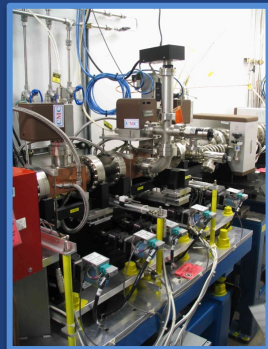




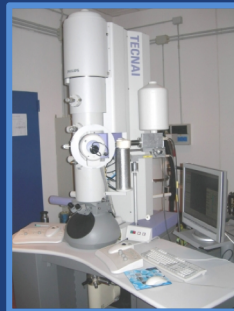
# Managing data from instruments



Next-Gen Sequencer



Advanced Light Source



Cryo-EM



MRI



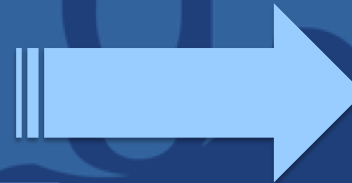
Light Sheet Microscope



Analysis store



High-durability, low-cost store



Remote visualization



Personal system





## Endpoints (Collections)

- **Software you deploy on your DTN – Globus Connect**
- **Storage abstraction**
  - All transfers happen between two endpoints
- **Collection  $\sim$  Endpoint**
- **Test / Demo Endpoints**
  - Globus Tutorial Endpoint 1/2
  - ESnet Read-Only \*
  - DME Datasets \*, DME PerfTest \*



# Globus Web App

File transfer

Data Sharing

Linked Identities

Groups



## Globus security

- **Access control**
  - Identities provided and managed by institution
  - Globus as identity broker: no access to institutional user credentials
  - Institution controls all access policies
- **Researchers can overlay sharing permissions**
  - Data remain at institution, not hosted by Globus
- **Automated integrity checks of transferred data**
- **High service availability**
- **Monitoring**
- **Encryption: all communications, data in transit (optional)**



## Globus for high assurance data management

- **Restricted data handling: PHI, PII, CUI**
- **Security controls: NIST 800-53, 800-171 Low**
- **Business Associate Agreement (BAA) w/UChicago**
  - University of Chicago has a BAA with Amazon
- **“Equivalent” UK/EU privacy contractual agreements**
  - e.g. to cover Data Processor requirements under GDPR



## High Assurance features

- **Additional authentication assurance**
  - Authenticate with a specific identity within a session
  - Reauthenticate after specified time period
- **Application instance isolation**
  - Authentication context is per application, per session
- **Forced encryption of data in transit**
- **Local audit logs (on data transfer nodes)**





## Globus security: Operational security

- **Intrusion detection and prevention**
- **Performance and health monitoring**
- **Logging**
- **Secure remote access, access control**
- **Uniform configuration management and change control**
- **Backups and disaster recovery**
- **Service data encrypted at rest**
- **AWS best practices (VPCs, IAM, Security Groups)**



# Protected Data Management

Authentication assurance

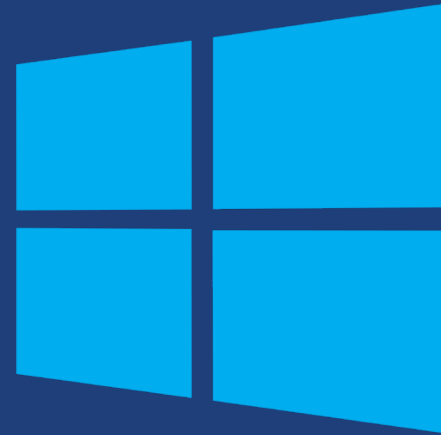
Session isolation



**...makes storage  
systems accessible  
via Globus**



## Globus Connect Personal



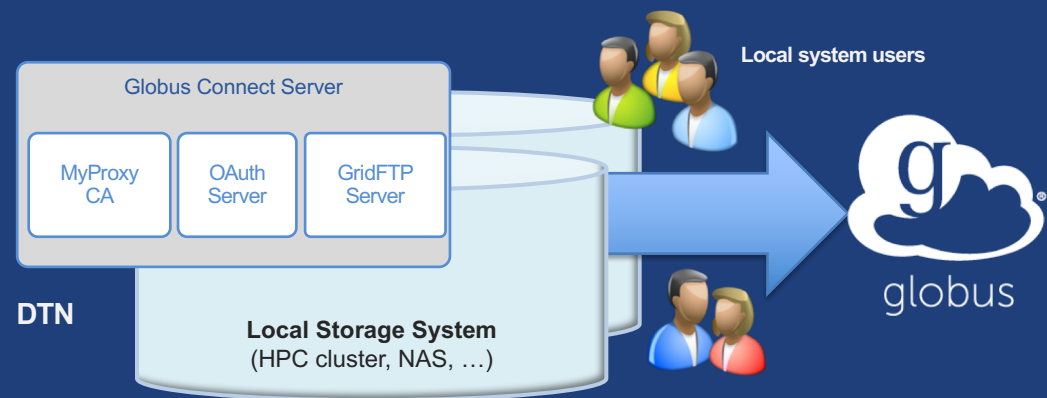
- **Rapid installation/removal by non-privileged account**
- **Zero configuration; auto updating**
- **Handles NATs**



# The Service Provider and System Administrator Perspective

# Globus Connect Server

- Makes your storage accessible via Globus
- Multi-user server, installed and managed by sysadmin
- Default access for all local accounts
- Native packaging  
Linux: DEB, RPM



[docs.globus.org/globus-connect-server-installation-guide/](https://docs.globus.org/globus-connect-server-installation-guide/)



# Storage Connectors - [globus.org/connectors](https://globus.org/connectors)

## Current

## Planned

IBM Spectrum Scale



ceph



Google Cloud

Lustre®



HPSS

Microsoft Azure



wasabi  
hot cloud storage

Western Digital



wasabi  
hot cloud storage



Which version of Globus Connect Server do I use?

**By default, assume you  
should use GCS v4**





## GCS: Common configuration options

- **Endpoints page**
  - Display Name
  - Visibility
  - Encryption
- **DTN configuration file**
  - RestrictPaths
  - Sharing
  - SharingRestrictPaths
  - IdentityMethod (MyProxy, CILogon, Oauth)

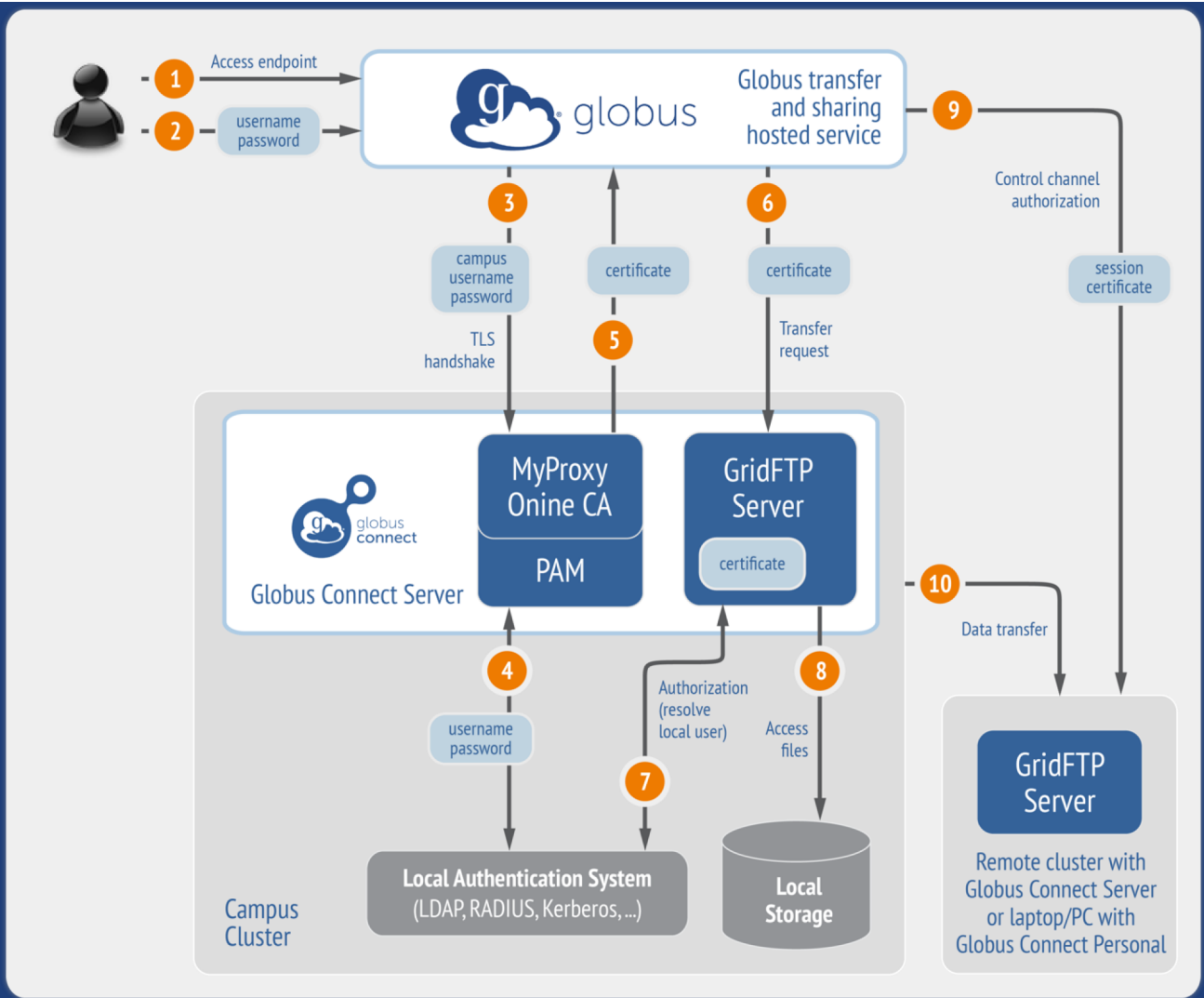


## Ports needed for Globus Connect Server v4

- **Inbound: 2811 (control channel)**
- **Inbound: 7512 (MyProxy), 443 (OAuth)**
- **Inbound: 50000-51000 (data channel)**
- **If restricting outbound connections, allow connections on:**
  - 80, 2223 (used during install/config)
  - 50000-51000 (GridFTP data channel)
- **GCSv5 removes some of these requirements – yay!**



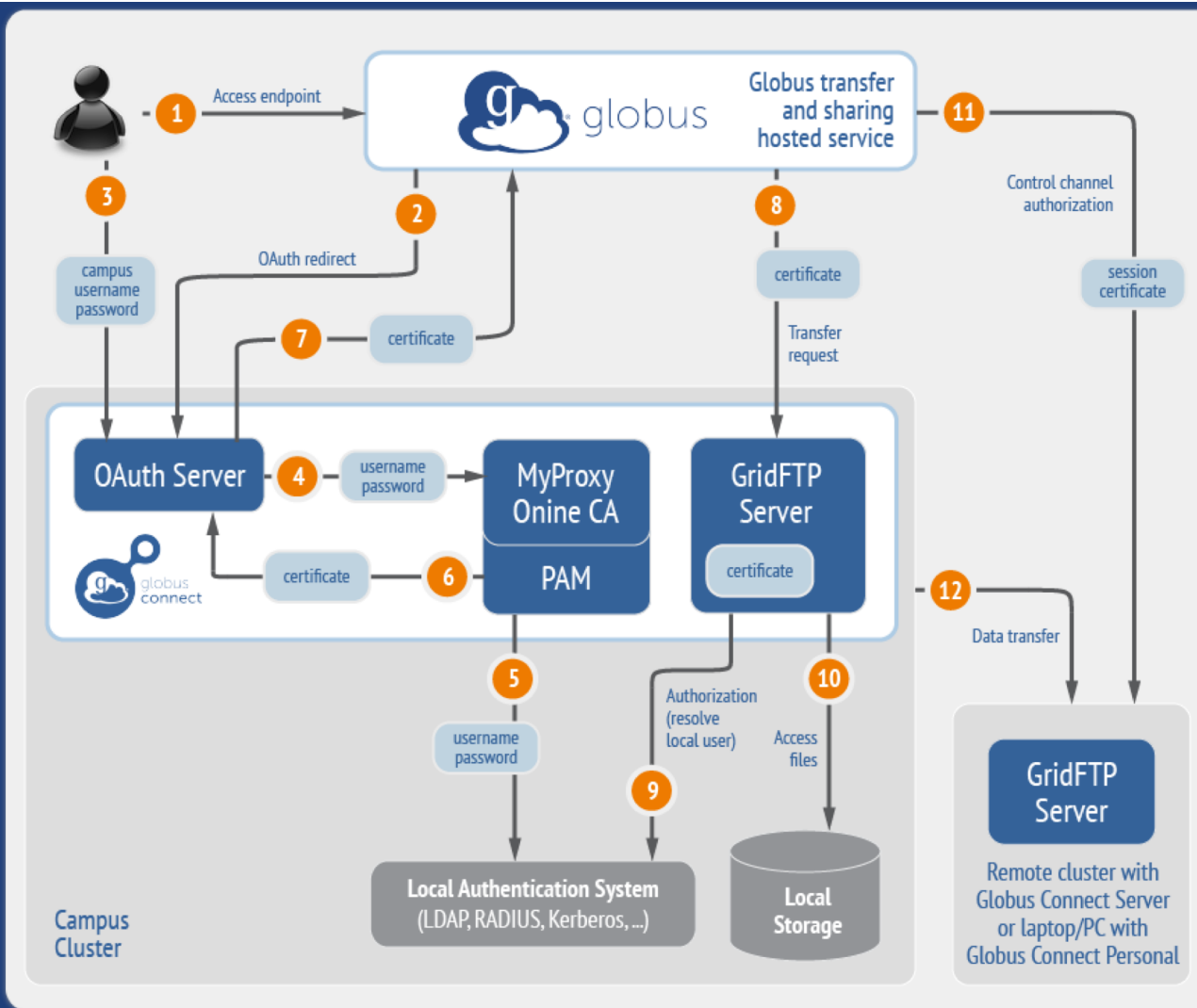
Endpoint activation using MyProxy



Default configuration  
*(avoid if at all possible)*



# Endpoint activation using MyProxy OAuth



Best practice configuration  
Just do it!  
Please...



## Alternative authentication methods and SSO

- **InCommon, EduGAIN, and other federations...**
  - Release R&S attributes to CILogon (especially ePPN)
  - Local account must match InCommon ID
  - In `/etc/globus-connect-server.conf` set:

```
AuthorizationMethod = CILogon
CILogonIdentityProvider = <institution_name>
```
  - Local account must match InCommon ID
- **Alternate identity providers**
  - Globus can add alternate IdPs to trusted list
  - Requires add-on subscription



## Subscription configuration

- **Subscription manager can create managed endpoints**
- **...Required for sharing, management console, usage reporting, ...**
- **Configurable via web app or CLI**



# Visibility and Control

Endpoint roles

Management console

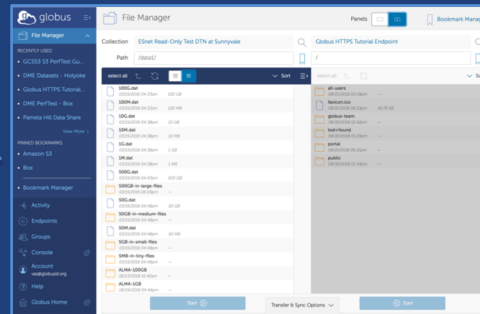
Usage reporting



# 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 (CLI)

- **Native application: [docs.globus.org/cli](https://docs.globus.org/cli)**
- **Open source, uses Python SDK**
- **`globus login` – get access and refresh tokens**
  - Tokens stored locally in `~/ .globus.cfg`
- **Service (transfer/auth) invocation uses tokens**
- **`globus logout` – delete tokens**

**[docs.globus.org/cli/examples](https://docs.globus.org/cli/examples)**



## UUIDs everywhere

- **UUIDs for endpoint, task, user identity, groups...**
- **Use search/list options**
- **get-identities for identity username to UUID**

```
$ globus endpoint search 'Globus Tutorial'  
$ globus task list  
$ globus get-identities vas@globus.org 14bf3755-  
6267-42f2-9e9c-ad324de4a1fb
```



## Batch Transfers

- Transfer tasks have one source/destination, but can have any number of files
- Provide input source-dest pairs via local file
- e.g. move files listed in files.txt from \$ep1 to \$ep2

```
$ ep1=e261ffb8-6d04-11e5-ba46-22000b92c6ec  
$ ep2=af7bda53-6d04-11e5-ba46-22000b92c6ec  
$ globus transfer $ep1:/share/godata/ $ep2:/~/ --  
batch --label 'CLI Batch' < files.txt
```



## Parsing CLI output

- **Default output is text; for JSON output use `--format json`**

```
$ globus endpoint search --filter-scope my-endpoints  
$ globus endpoint search --filter-scope my-endpoints --  
format json
```

- **Extract specific attributes using `--jmespath <expression>`**

```
$ globus endpoint search --filter-scope my-endpoints --  
jmespath 'DATA[].[id, display_name]'
```



## Permission management

- **Set and manage permissions on shared endpoint**
- **Requires access manager role**

```
$ share=<shared_endpoint_UUID>
$ globus endpoint permission create --permissions r --
identity tuecke@globus.org $share:/NCARTest/
$ globus endpoint permission list $share
$ globus endpoint permission delete $share <perm_UUID>
```



## Automation Examples

- **Syncing a directory**
  - bash script; calls the Globus CLI
  - Python module; run as script or import as module
- **Staging data in a shared directory**
  - bash and Python variants
- **Removing directories after files are transferred**
  - Python script

[github.com/globus/automation-examples](https://github.com/globus/automation-examples)



# The (Portal/Gateway/App/...) Developer Perspective



How can I (more tightly)  
integrate Globus into my  
research workflows?





Globus serves as...

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



# Example: Data repositories

The screenshot displays the NCAR Research Data Archive interface. The main header includes the NCAR logo and the text "Research Data Archive Computational & Information Systems Lab". Below this is a navigation bar with buttons for "Home", "Find Data", "Ancillary Services", "About/Contact", "Data Citation", "Web Services", and "For Staff". The main content area is titled "NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products" with the identifier "ds094.2". A "Data Selection Summary" modal window is open, showing a list of selected files and options for data delivery and format. Below the modal, a table provides detailed information about the data products.

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
<b>Union of Available Products</b>	Web File Listing	Request Globus Invitation	Get a Subset	TDS Access	GLADE File Listing	HPSS File Listing
Diurnal monthly means	Web File Listing		Get a Subset		GLADE File Listing	HPSS File Listing



# Example: Analysis services

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 in the details below to **register** and/or **phasing job**. If you need more information, see the [about](#) page.

Full name

Organisation

Email address

What is this? [+](#)  
Globus user identity

[Next](#)

News [@sangerimpute](#)

## DLHub

Data and Learning Hub for Science

A simple way to find, share, publish, and run machine learning models and discover training data for science

### Documentation

Examples


Read the Docs

Python SDK

CLI

### Papers and Presentations

 DLHub on ArXiv

 DLHub Slides





## Example: Instruments

### DMagic a Globus implementation at the APS



<http://dmagic.readthedocs.org>



Tomopy

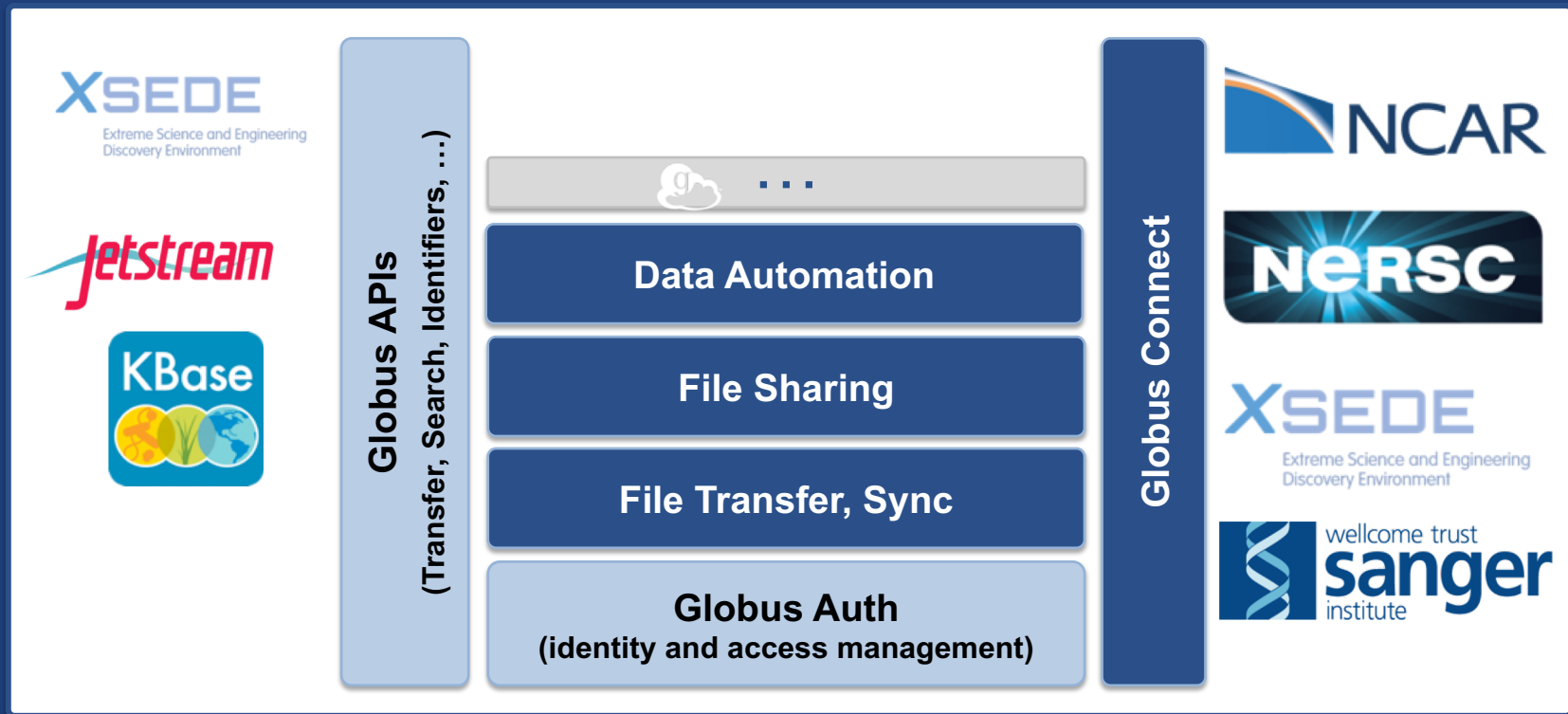
Tomographic  
reconstruction in Python

Doga Gursoy

<http://tompy.readthedocs.org>



# Globus Platform-as-a-Service





## Globus Auth addresses security challenges

- **Make it easy for developers to provide login for their apps (web, mobile, desktop, command line)**
- **...and protect all REST API communications**
  - App → Globus service (MRDP, Jupyter Notebook)
  - App → non-Globus service (graph service in MRDP)
  - Service → Service
- **...while**
  - Not introducing yet another identity
  - Providing a platform to consolidate existing identities
  - Providing a least privileges security model (via consents)
  - Being web friendly and language/framework agnostic



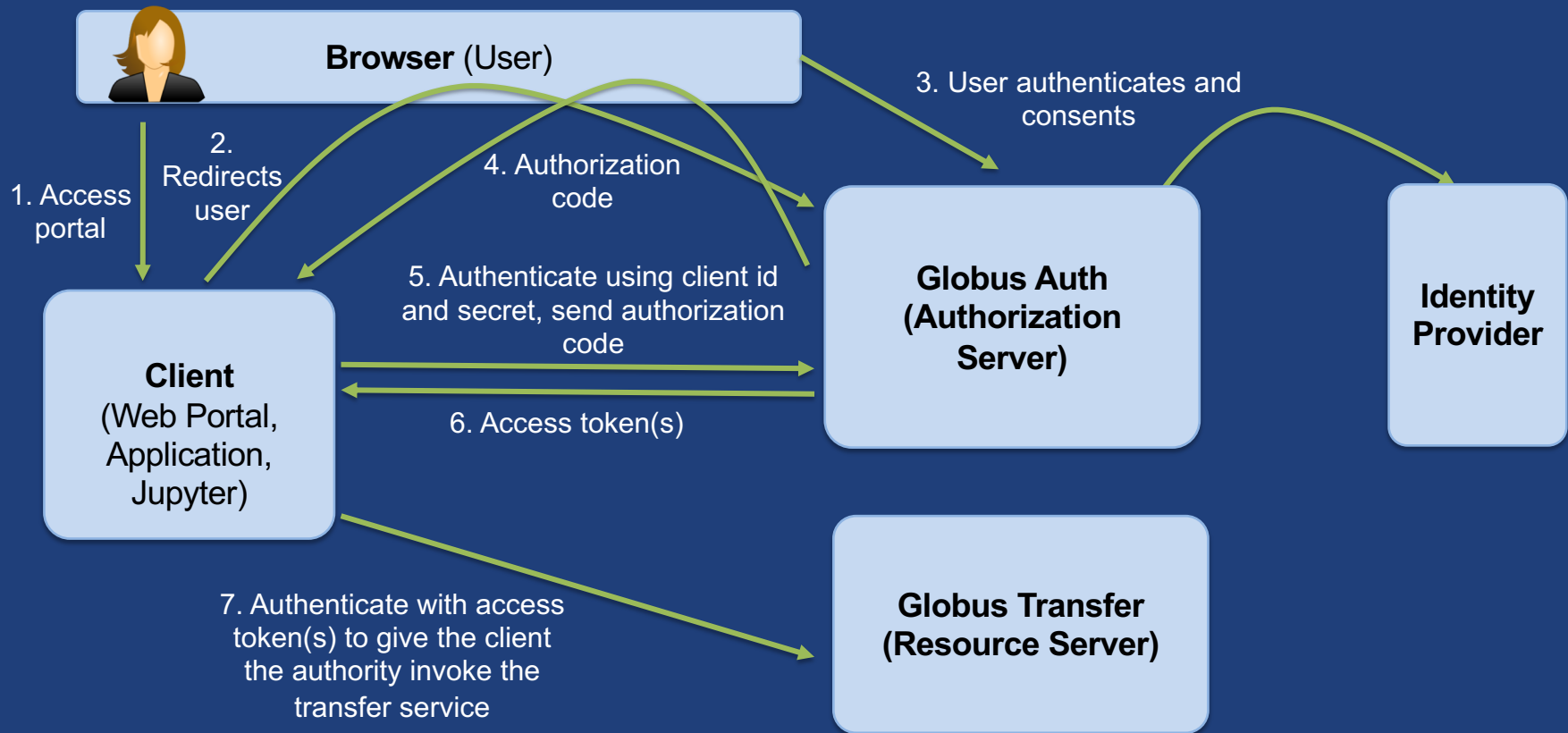
Based on widely used web standards

- **OAuth 2.0 Authorization Framework (a.k.a. OAuth2)**
- **OpenID Connect Core 1.0 (a.k.a. OIDC)**
- **Access via OAuth2 and OIDC libraries of your choice**
  - Google OAuth Client Libraries, Apache mod\_auth\_openidc, etc.
  - Globus Python SDK

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



# Auth Example: Authorization Code Grant







## Globus Transfer API

- **Globus Web App consumes public Transfer API**
- **Resource named by URL (standard REST approach)**
  - Query params allow refinement (e.g., subset of fields)
- **Globus APIs use JSON for documents and resource representations**
- **Requests authorized via OAuth2 access token**
  - Authorization: Bearer asdfkqhafsdafeawk

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



## Globus Python SDK

- **Python client library for the Globus Auth and Transfer REST APIs**
- **`globus_sdk.TransferClient` class handles connection management, security, framing, marshaling**

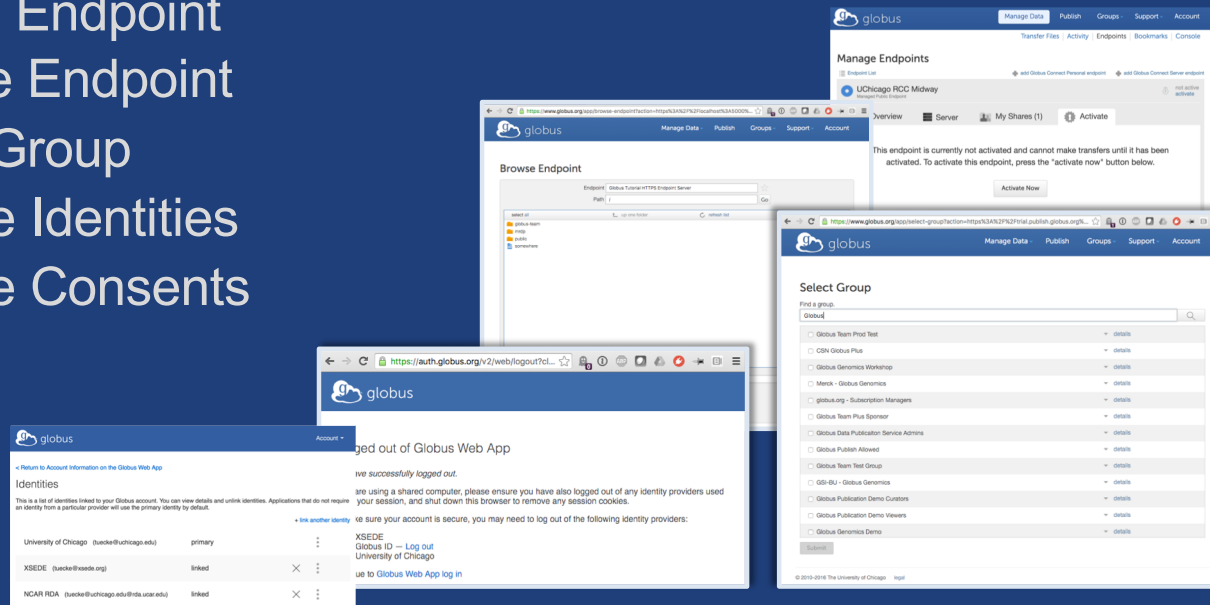
```
from globus_sdk import TransferClient  
tc = TransferClient()
```

**[globus.github.io/globus-sdk-python](https://globus.github.io/globus-sdk-python)**



# Globus Helper Pages

- Globus pages designed for use by your web apps
  - Browse Endpoint
  - Activate Endpoint
  - Select Group
  - Manage Identities
  - Manage Consents
  - Logout



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



# 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
- List

## Python SDK

Installation

The Globus SDK requires Python 2.6+ or 3.2+. If a su...

The simplest way to install the Globus SDK is using th...

```
pip install globus-sdk
```

This will install the Globus SDK and it's dependencies...

Bleeding edge versions of the Globus SDK can be in...

```
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-63e...>
- Installed Globus Python SDK

## Jupyter Notebook

```
In [15]: from __future__ import print
tutorial_endpoint_1 = "ddb59ae1-0d04-11e5-ba46-22000b92c6ec" # endpoint "Glo
tutorial_endpoint_2 = "ddb59af0-6d04-11e5-ba46-22000b92c6ec" # endpoint "Glo
tutorial_users_group = "50b6a29c-63ac-11e4-8062-22000ab68755" # group "Tutori
```

## Configuration

First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you c...

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)

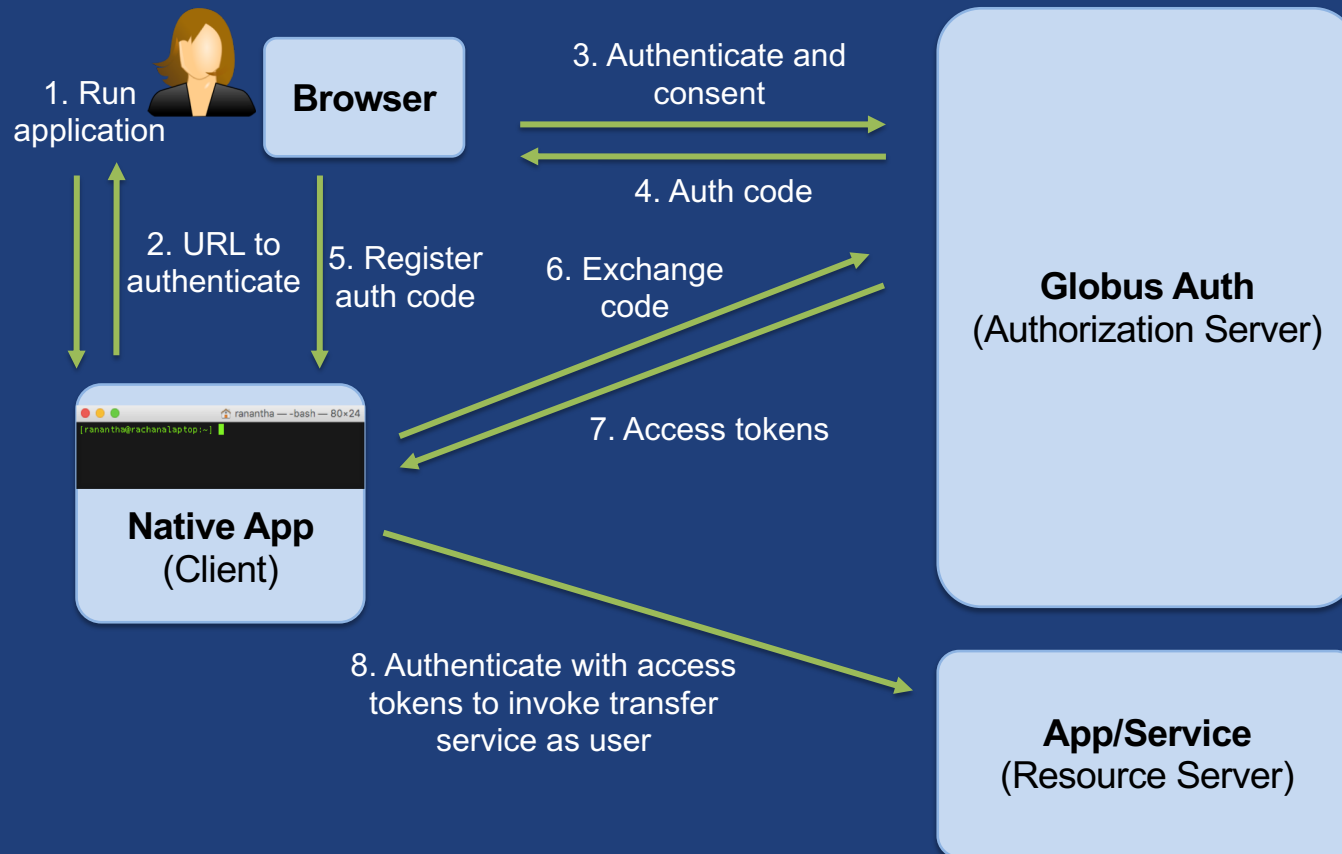


## Globus Auth: Native apps

- **Client that cannot keep a secret, e.g...**
  - Command line, desktop apps
  - Mobile apps
  - Jupyter notebooks
- **Native app is registered with Globus Auth**
  - Not a confidential client
- **Native App Grant is used**
  - Variation on the Authorization Code Grant
- **Globus SDK:**
  - To get tokens: `NativeAppAuthClient`
  - To use tokens: `AccessTokenAuthorizer`



# Native App grant



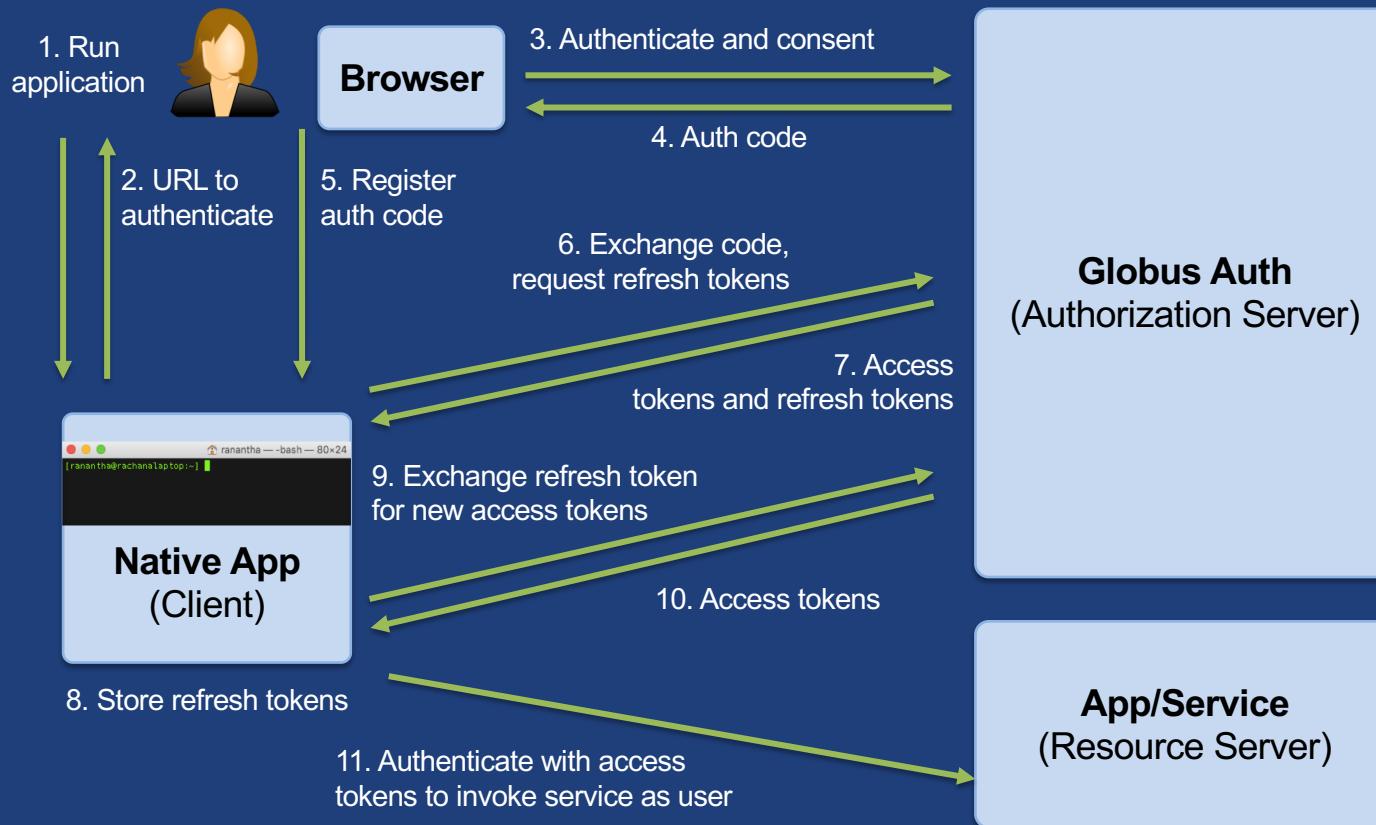


# Refresh tokens

- **Common use cases**
  - Portal checking transfer status when user is not logged in
  - Running command line app from script
- **Refresh tokens issued to client, in particular scope**
- **Client uses refresh token to get access token**
  - Confidential client: `client_id` and `client_secret` required
  - Native app: `client_secret` not required
- **Refresh token good for 6 months after last use**
- **Consent rescindment revokes all tokens**



# Refresh tokens







## Native App/Refresh Tokens Sample Code

[github.com/globus/native-app-examples](https://github.com/globus/native-app-examples)

- `./example_copy_paste.py`
  - User copies and pastes code to the app
- `./example_copy_paste_refresh_token.py`
  - Stores refresh token locally, uses it to get new access tokens
- **See README for installation**



# Automating data flows at scale



## Instrument Use Cases

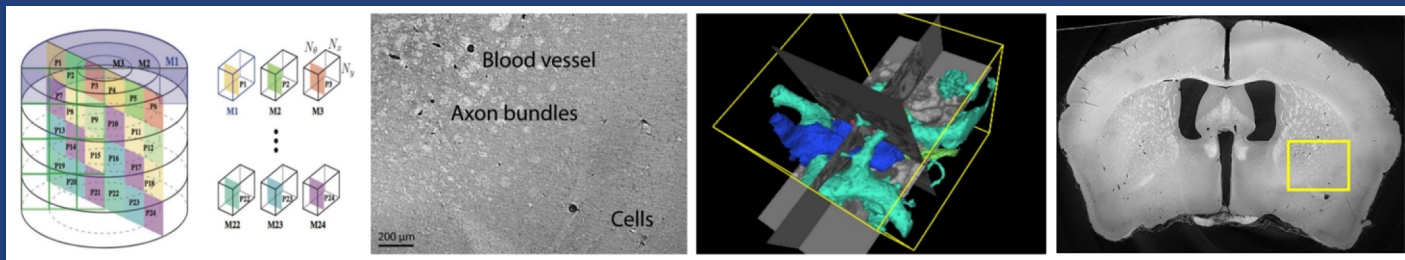
- **Advanced Photon Source**
  - Connectomics
  - Time series spectroscopy
- **Scanning Electron Microscope**
  - Materials science
- **Cryo-electron Microscope**

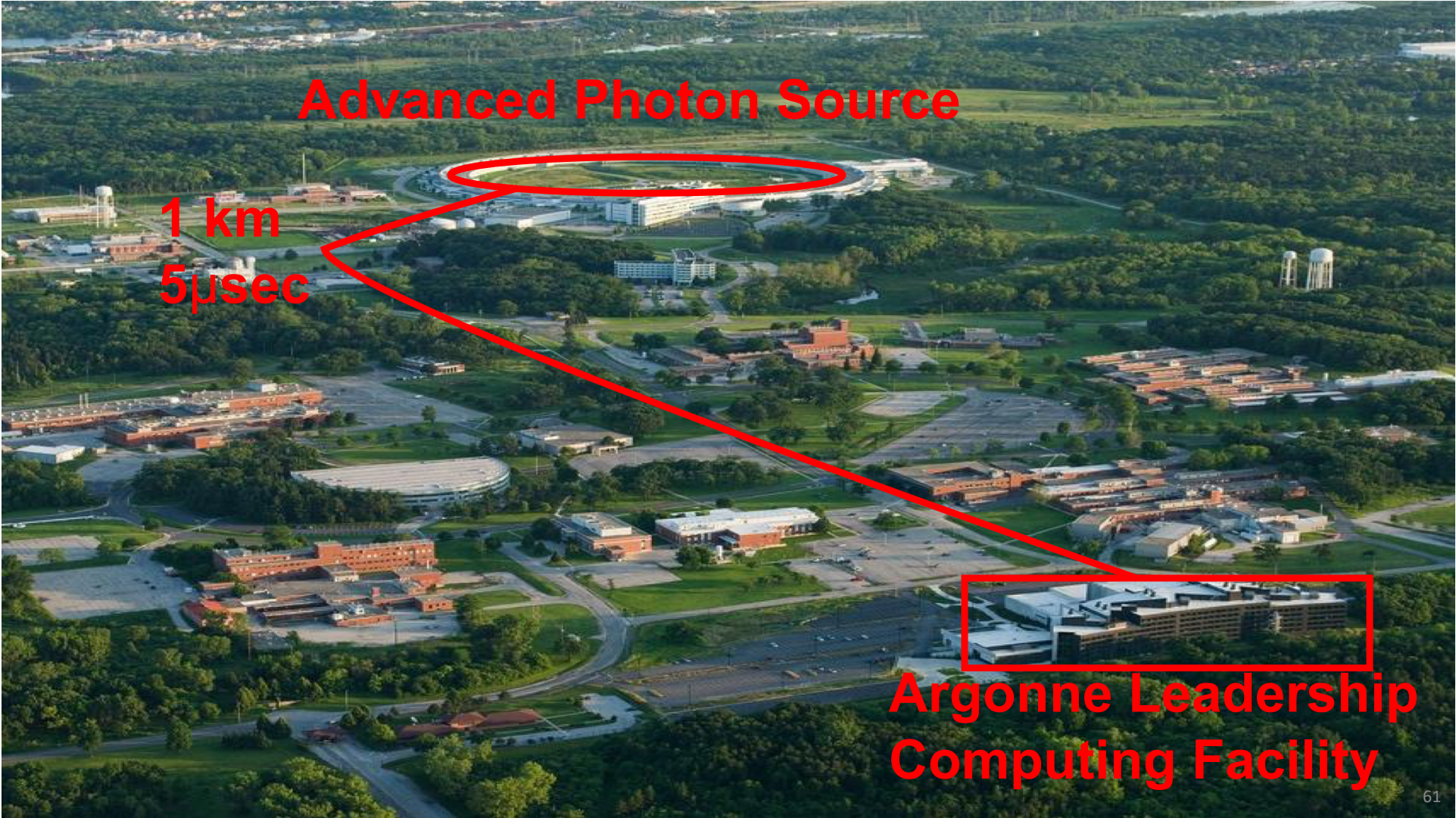




# UChicago Kasthuri Lab: Brain aging and disease

- Construct connectomes—mapping of neuron connections
- Use APS synchrotron to rapidly image brains
  - Beam time available once every few months
  - ~20GB/minute for large (cm) unsectioned brains
- Generate segmented datasets/visualizations for the community
- Perform semi-standard reconstruction on all data across HPC resources





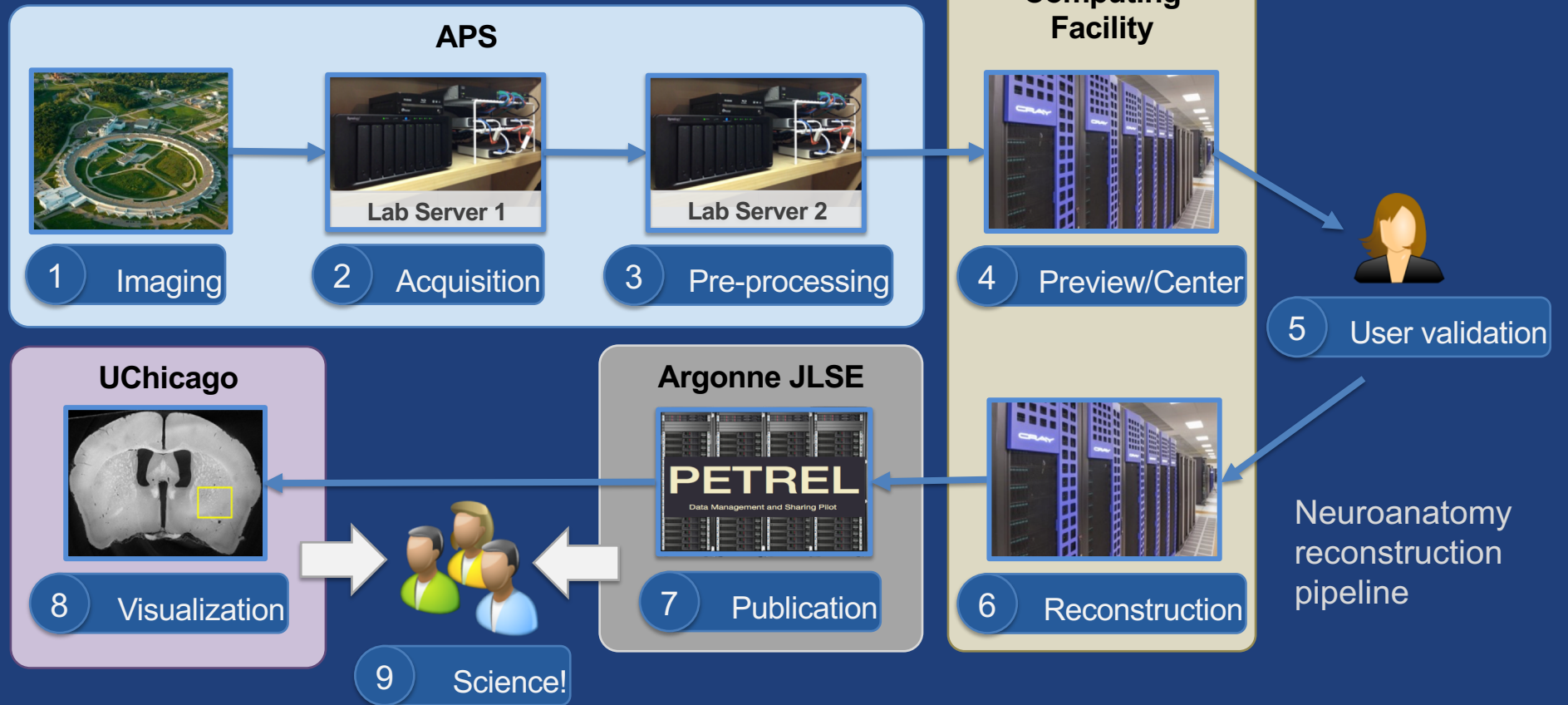
**Advanced Photon Source**

**1 km**  
**5 μsec**

**Argonne Leadership  
Computing Facility**



# Building the connectome





Data Example

# ALCF Data Discovery Portal

<https://petreldata.net>



Prototypical Globus Platform example:

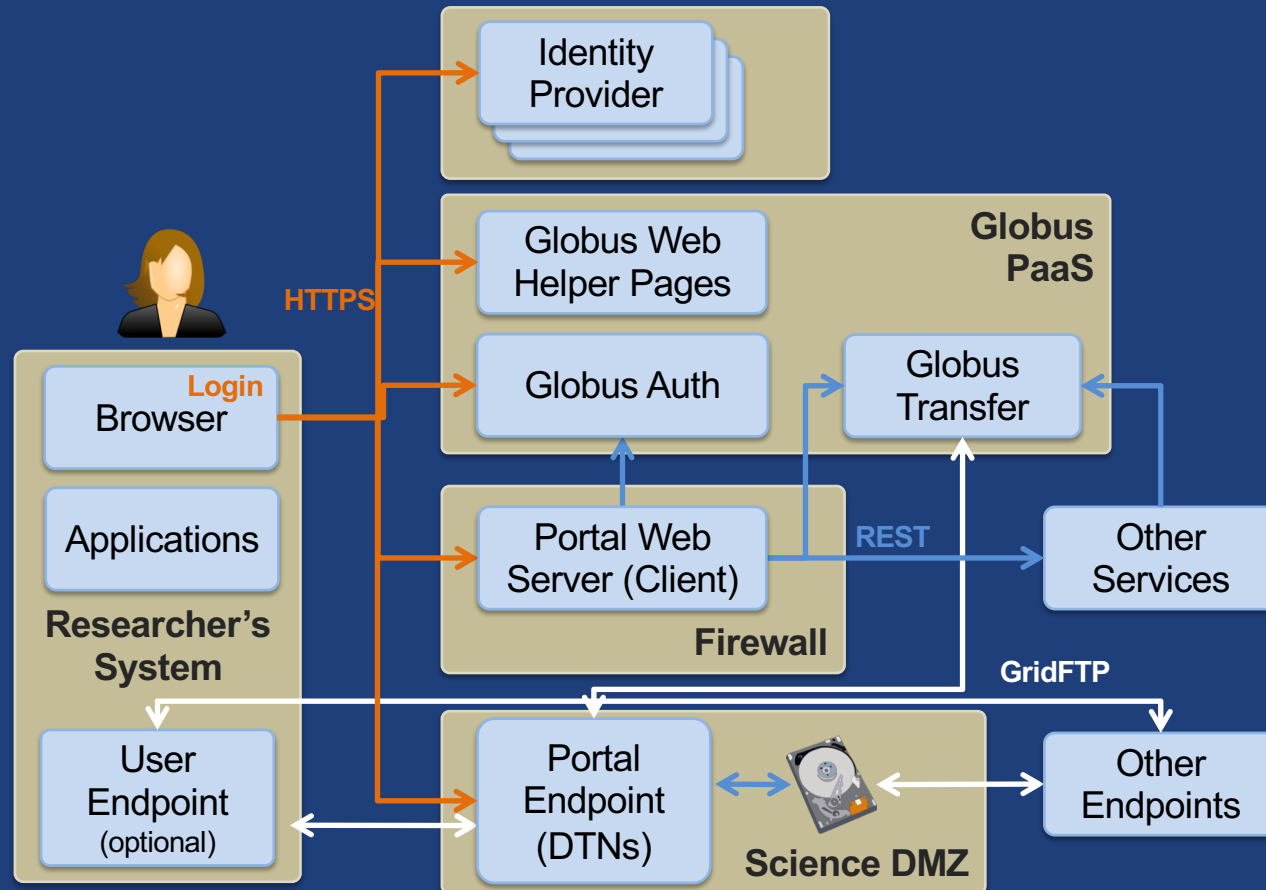
# Modern Research Data Portal

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



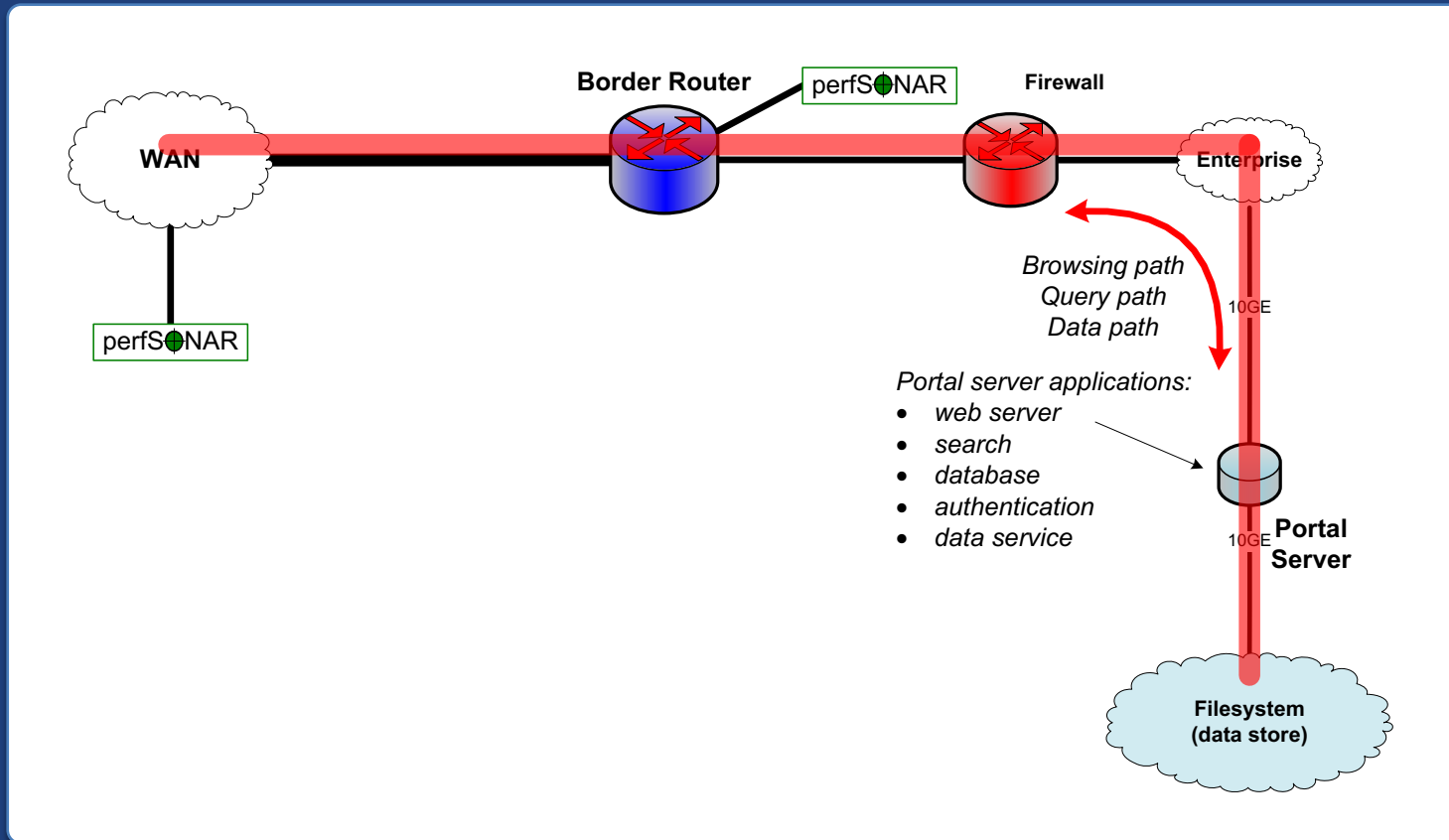


# Modern Research Data Portal Design Pattern



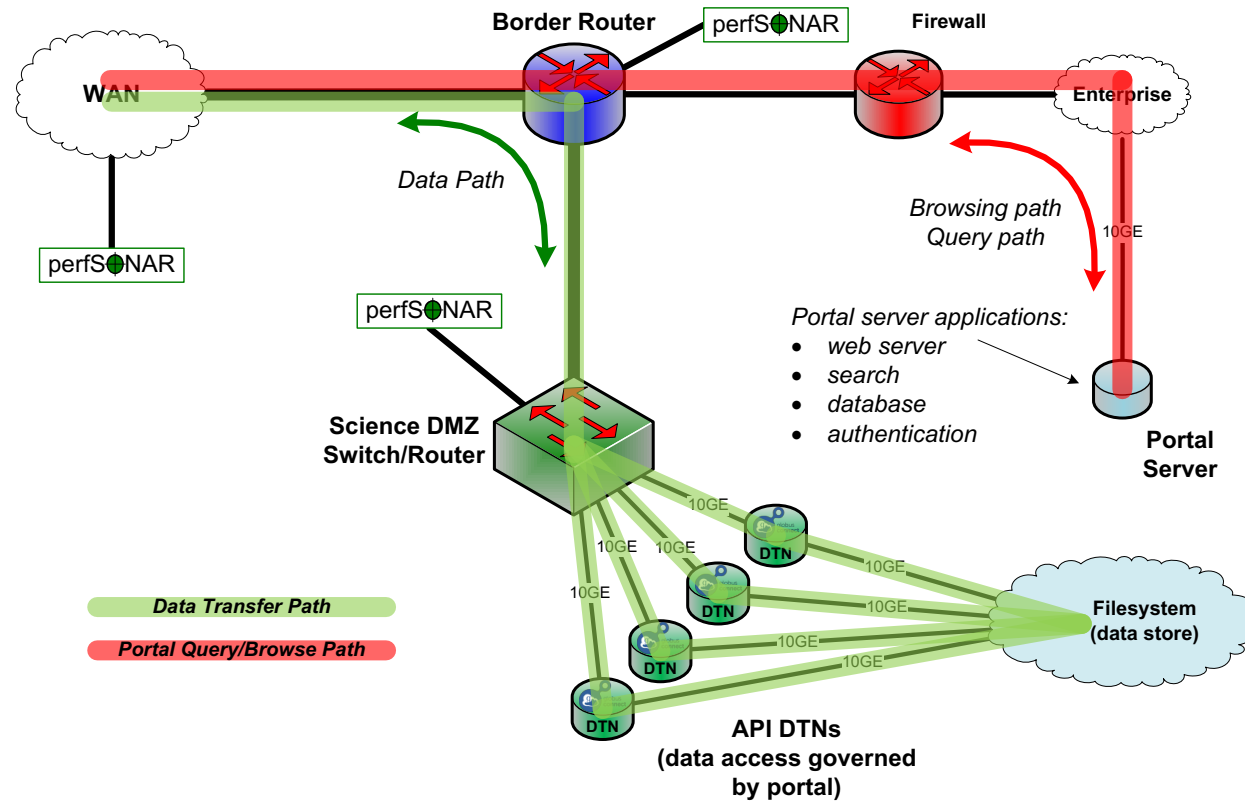


# Legacy research data portal architecture





# MRDP Architecture





## Relevant data sharing elements

- **One-time, manual creation of shared endpoint**
- **Permissions set per folder on shared endpoint**
- **Permissions management can be automated**
  - User: researcher@uchicago.edu
  - Group: search for group to get Group UUID
  - Application ...yes, apps are people too!
- **Roles for management of endpoint and tasks**
  - Grant rights to other users, groups or applications
- **Access manager role for managing permissions**

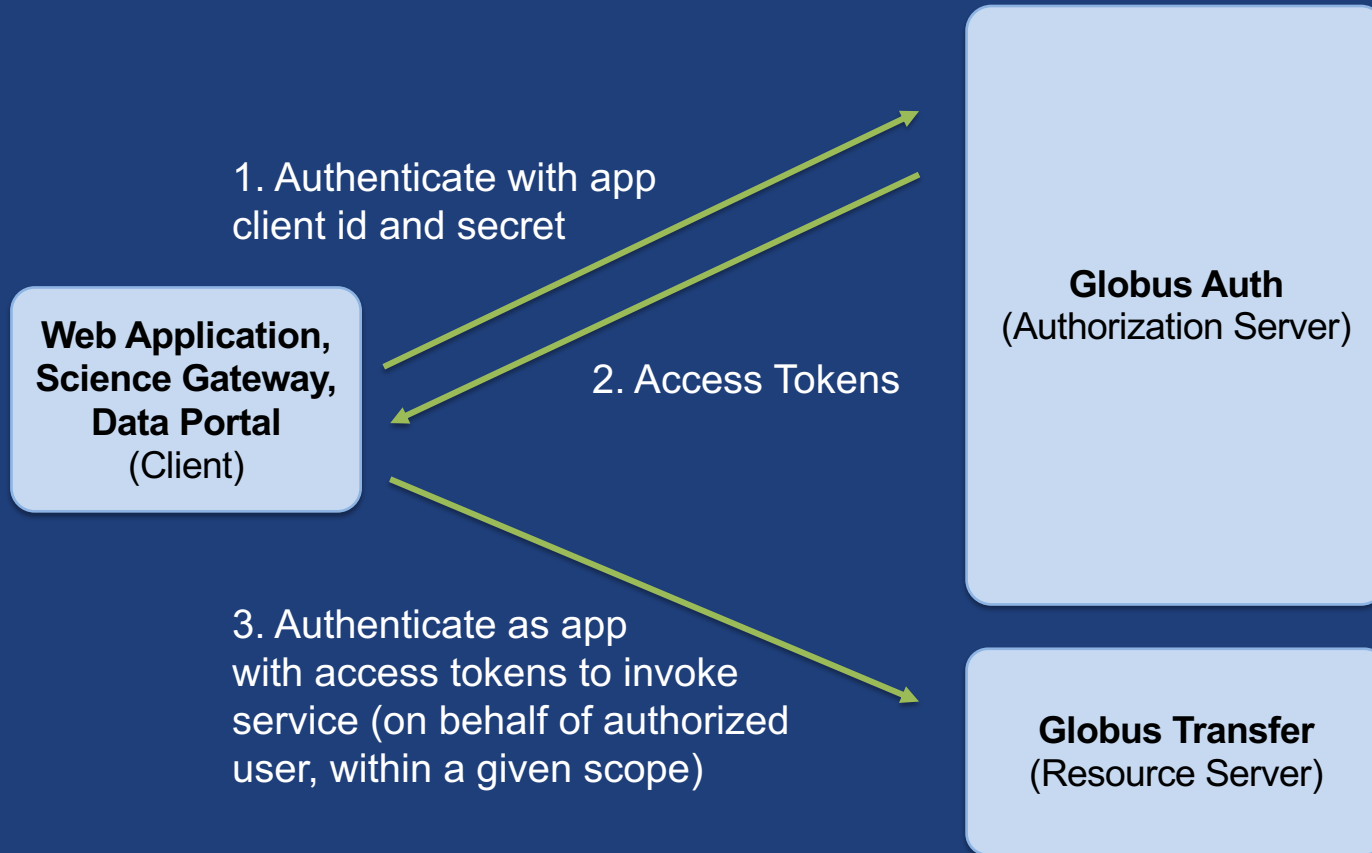


## Security elements: Confidential app

- **Uses client id and secret**
- **Ensure application is on a secure device**
- **Set up policy for rotation of secret (limited life tokens)**
- **Identity: <app\_client\_id>@clients.auth.globus.org**



# Client credential grant





## Confidential clients must be registered

- **Register at [developers.globus.org](https://developers.globus.org)**
  - Redirects; e.g. `https://your_fqdn/app/authcallback`
  - Scopes
    - e.g. `globus:auth:scope:transfer.api.globus.org:all`
    - e.g. `profile, email, openid`
- **Get client id and secret**
- **Ensure secret is properly protected, rotated**



## Support resources

- **Globus documentation:** [docs.globus.org](https://docs.globus.org)
- **Sample code:** [github.com/globus](https://github.com/globus)
- **Helpdesk and issue escalation:** [support@globus.org](mailto:support@globus.org)
- **Customer engagement team**
- **Globus professional services team**
  - Assist with portal/gateway/app architecture and design
  - Develop custom applications that leverage the Globus platform
  - Advise on customized deployment and integration scenarios





## 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)