Introduction to Globus: SaaS for Research Data Management

Vas Vasiliadis
vas@uchicago.edu

Rachana Ananthakrishnan
rachana@globus.org

Penn State University — June 29, 2017

Research data management today

Index?

How do we...

...move?

...share?

...discover?

...reproduce?

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 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

Pre-processed Data

Analysis store

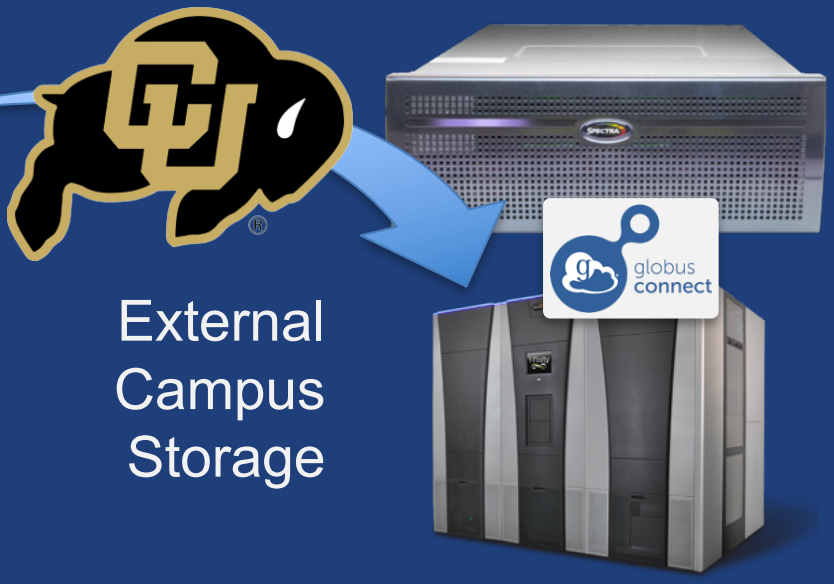
Raw
Source
Data

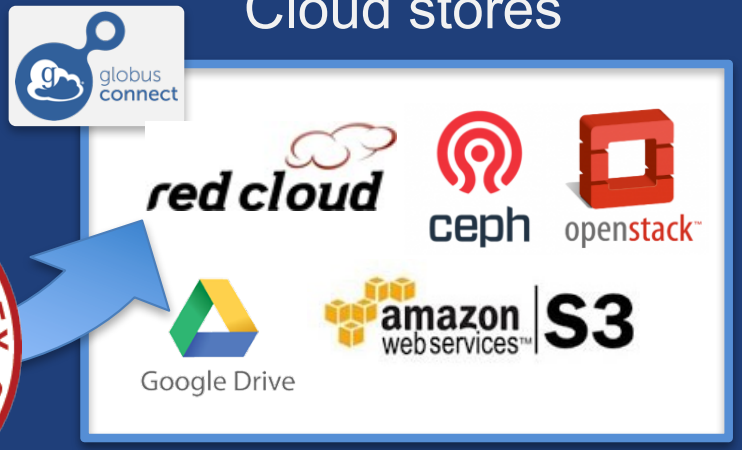
High durability,
low cost store

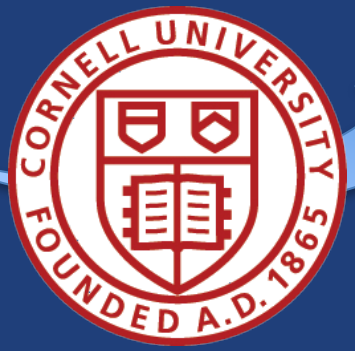
globus
connect

Amazon
Glacier

Bridge to collaborators







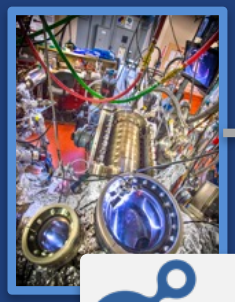
Bridge to community/public

Project Repositories,
Replication Stores

Public Repositories

Globus and the research data lifecycle

Instrument



Globus transfers files reliably, securely

2

Compute Facility

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

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

Transfer

3

Share

1

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

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

6

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

Publication Repository

5

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

Publish

6

8

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

Personal Computer

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

Discover

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

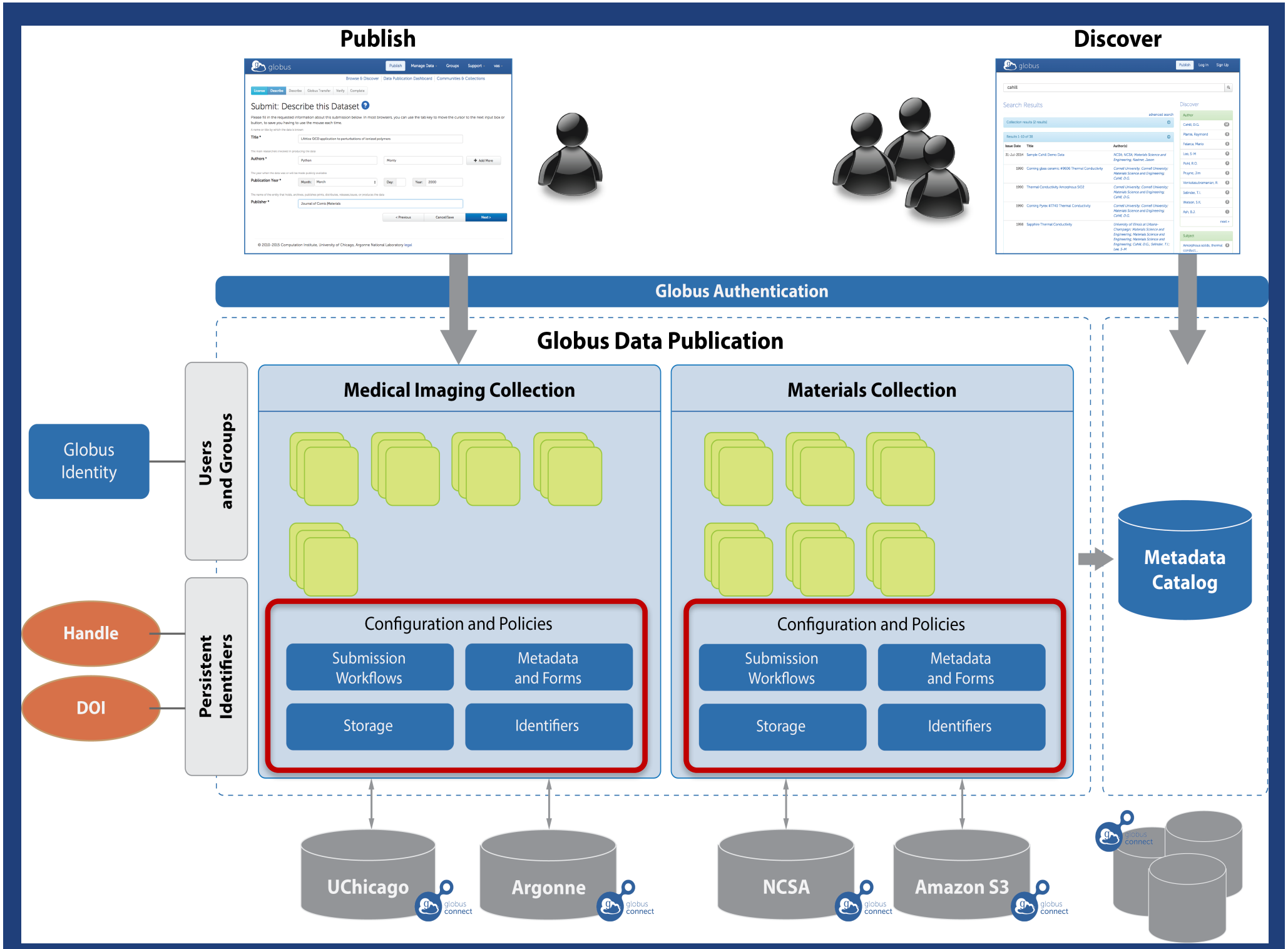
File Transfer

File Sharing

Federated Identity

Data Publication and Discovery

https://publish.globus.org


Publish
Discover
Globus Authentication
Globus Data Publication
Globus Identity
Users and Groups
Handle
DOI
Persistent Identifiers
Medical Imaging Collection
Materials Collection
Configuration and Policies
Submission Workflows
Metadata and Forms
Storage
Identifiers
Metadata Catalog
UChicago
Argonne
NCSA
Amazon S3
globus connect

Peer reviewed paper data

(Re)format…

- PDF/A
- HDF
- …

Describe…

- Dublin core metadata
- Domain metadata
- Provenance info
- ...

Community and public repositories

Persistent identifier

- DOI

Curation workflow

- Review → Update → Resubmit cycle

Demonstration
Data Publication

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**

Moving data between your laptop and another system

Exercise: Log in & transfer files

1. Go to: 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 an ESnet test endpoint and your laptop

Sharing Data

Share files

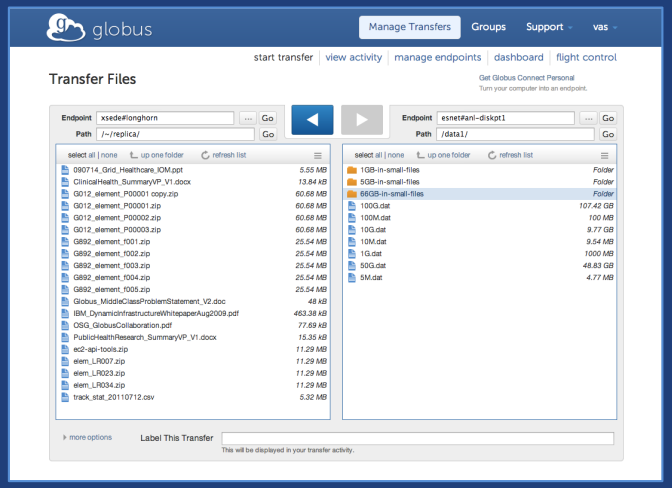
1. **Join the “Tutorial Users” groups**
   - Go to “Groups”, search for “tutorial”
   - Select group from list, click “Join Group”
2. **Create a shared endpoint on your laptop**
3. **Grant your neighbor permissions on your shared endpoint**
4. **Access your neighbor’s shared endpoint**

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



Web

Globus service

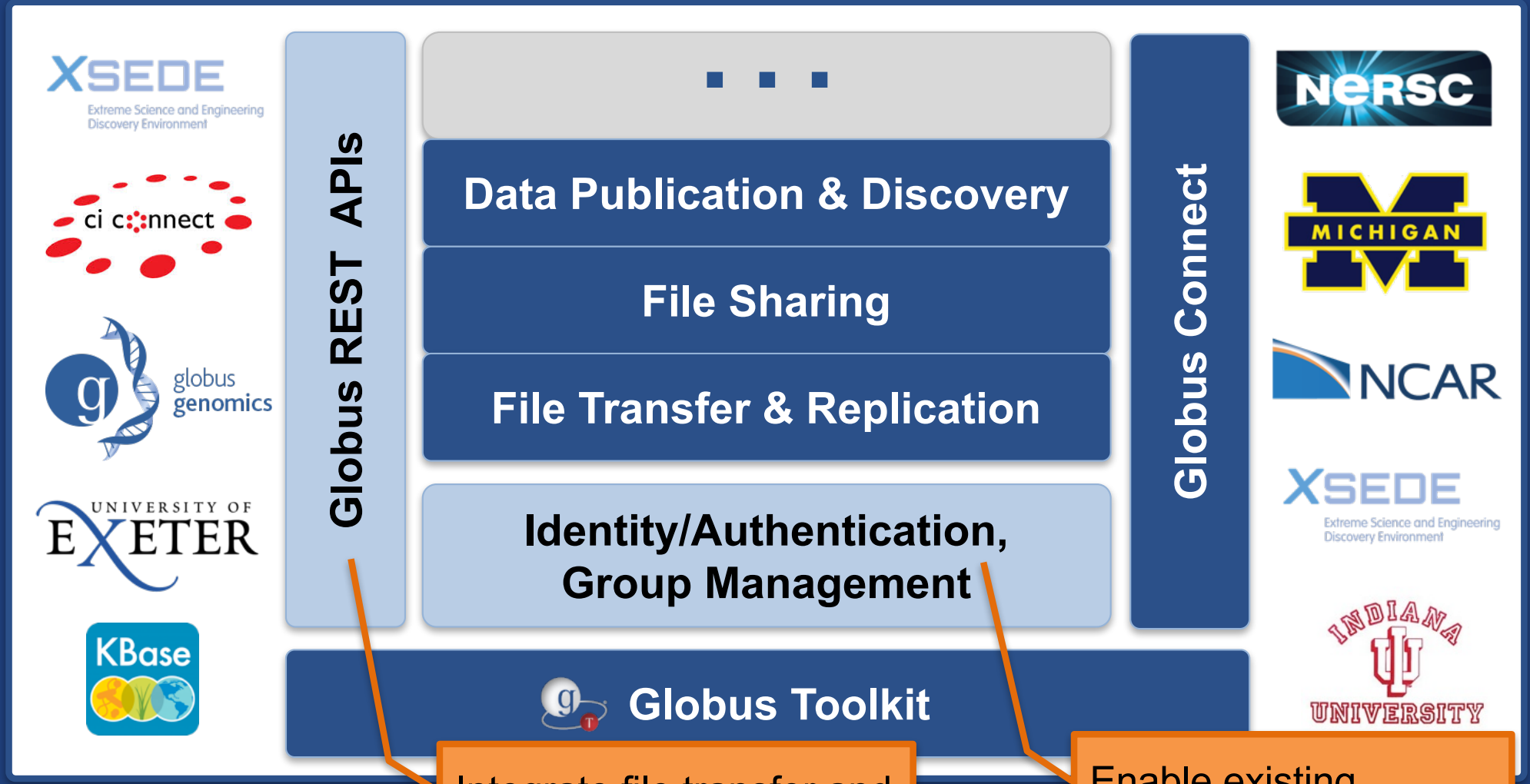
```
(globus-cli) jupiter:~ vas$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

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

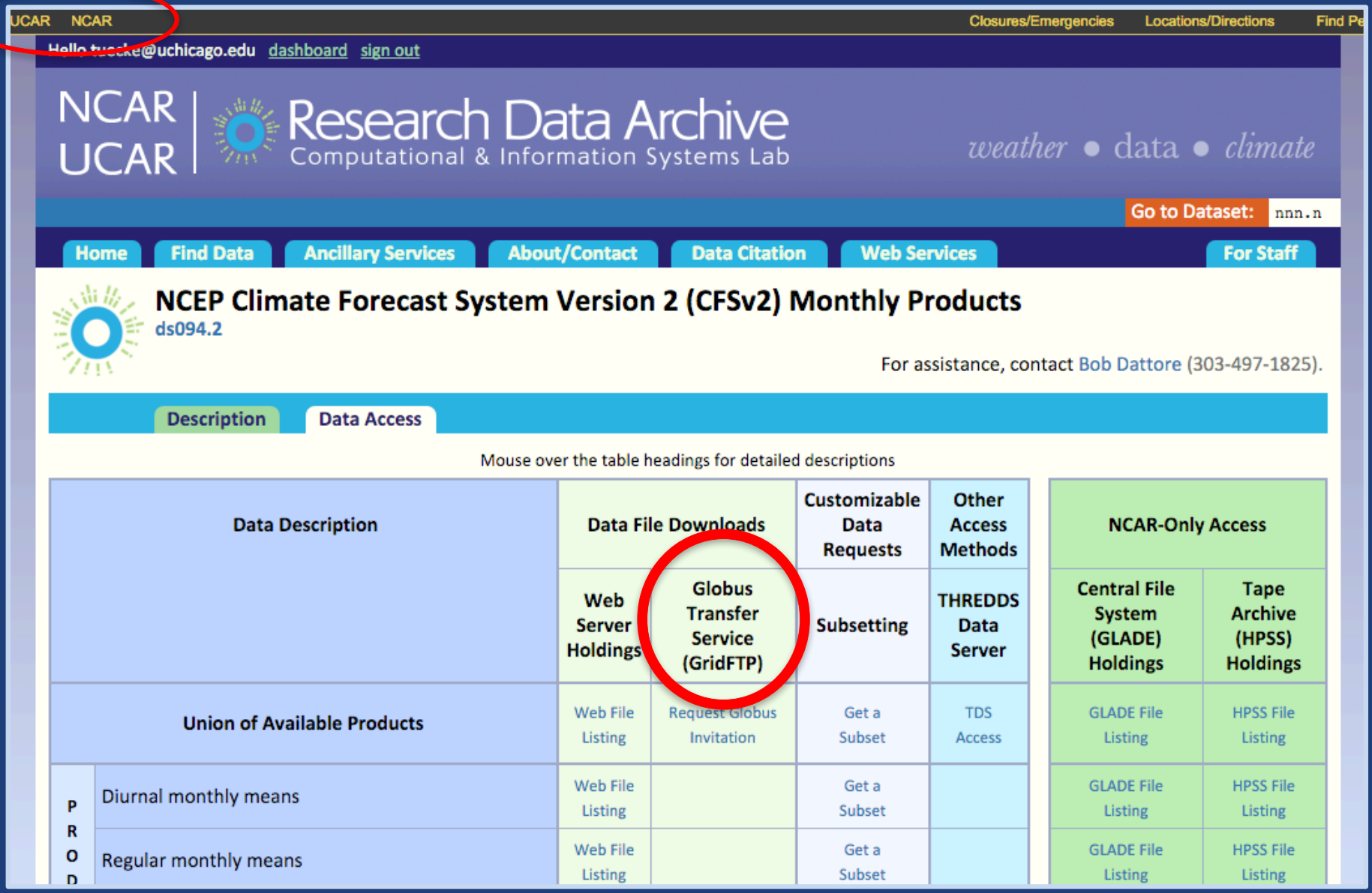
```
GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
...
```

Rest API

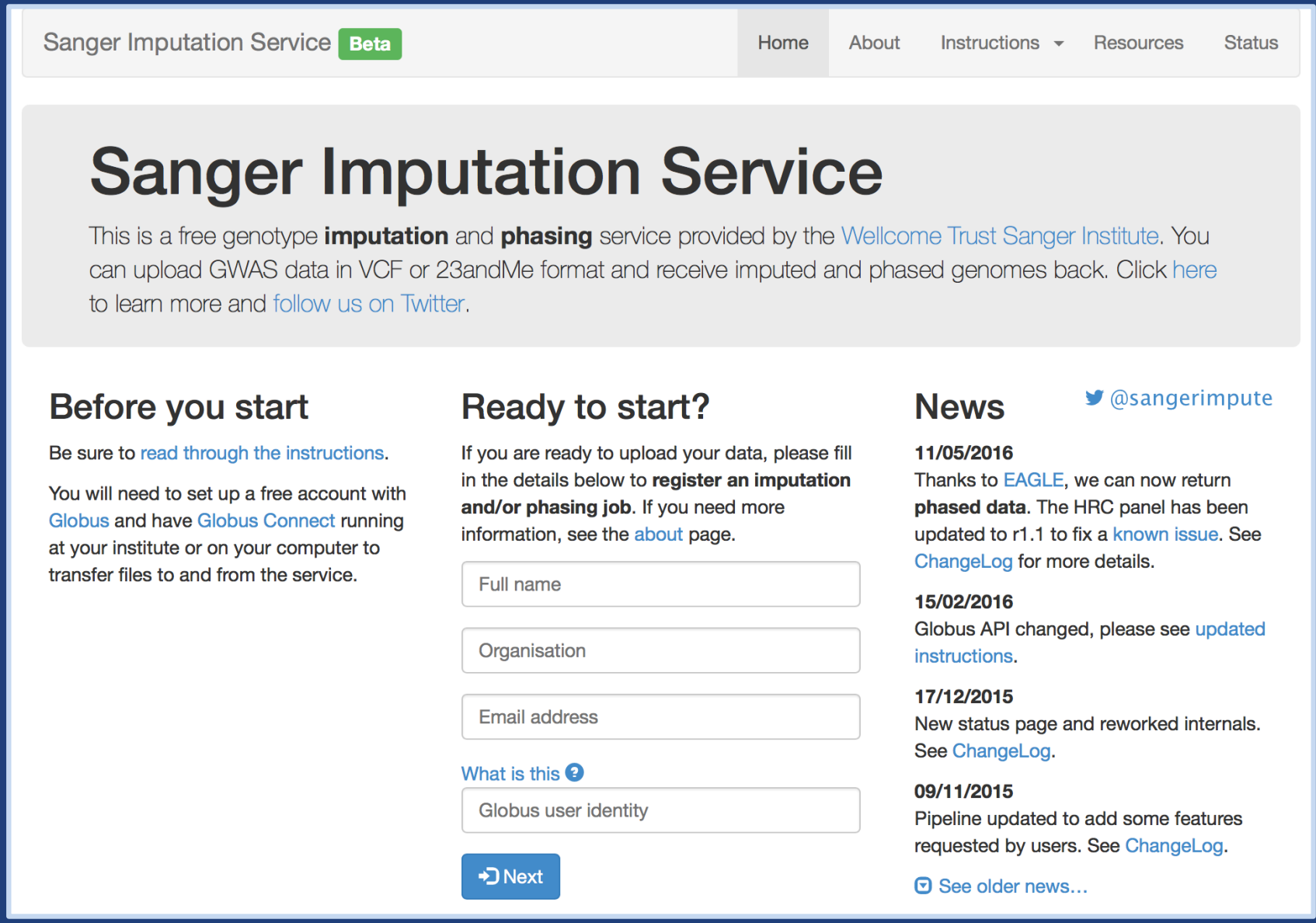
Globus as PaaS



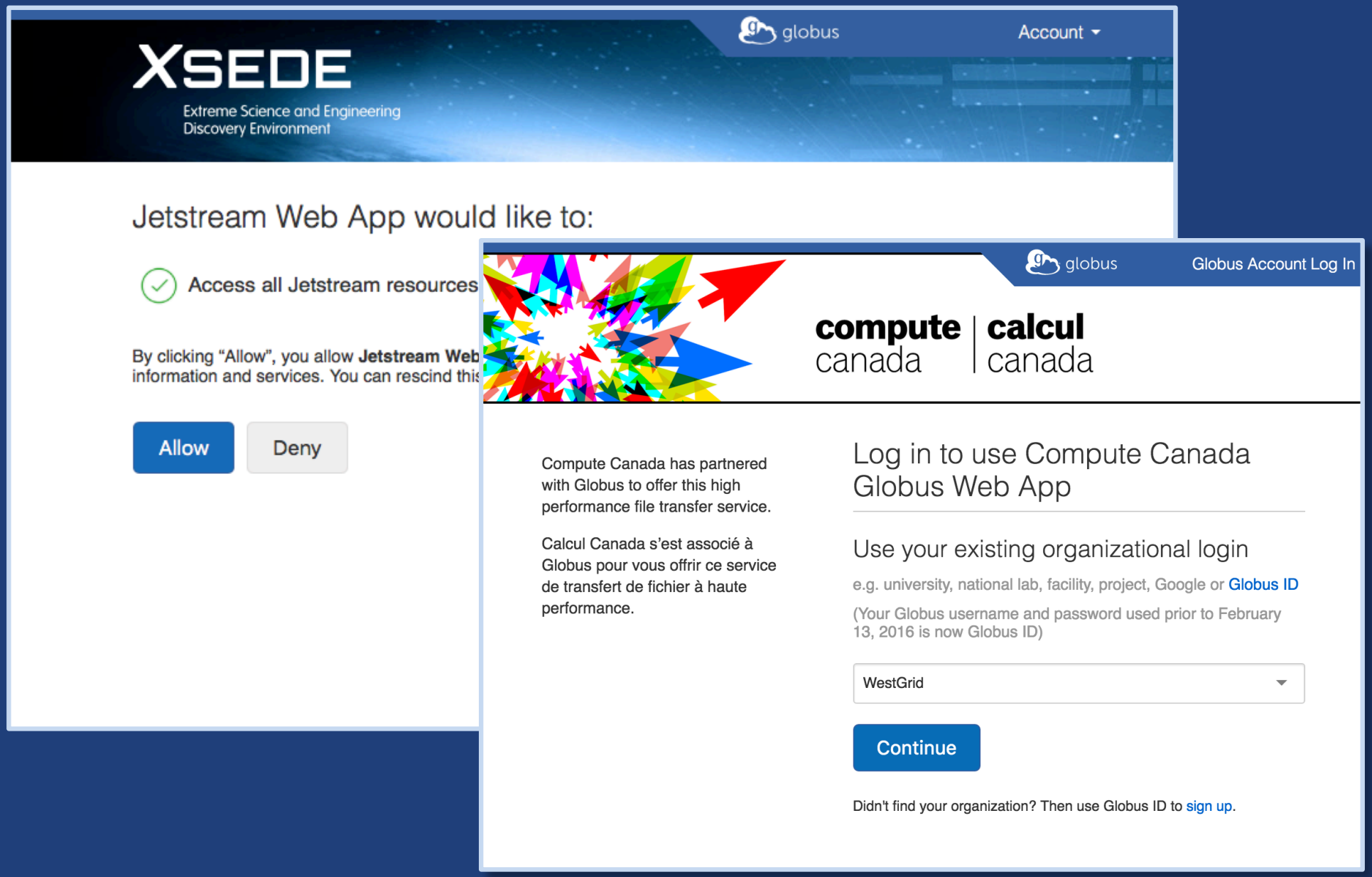
Data App: NCAR RDA



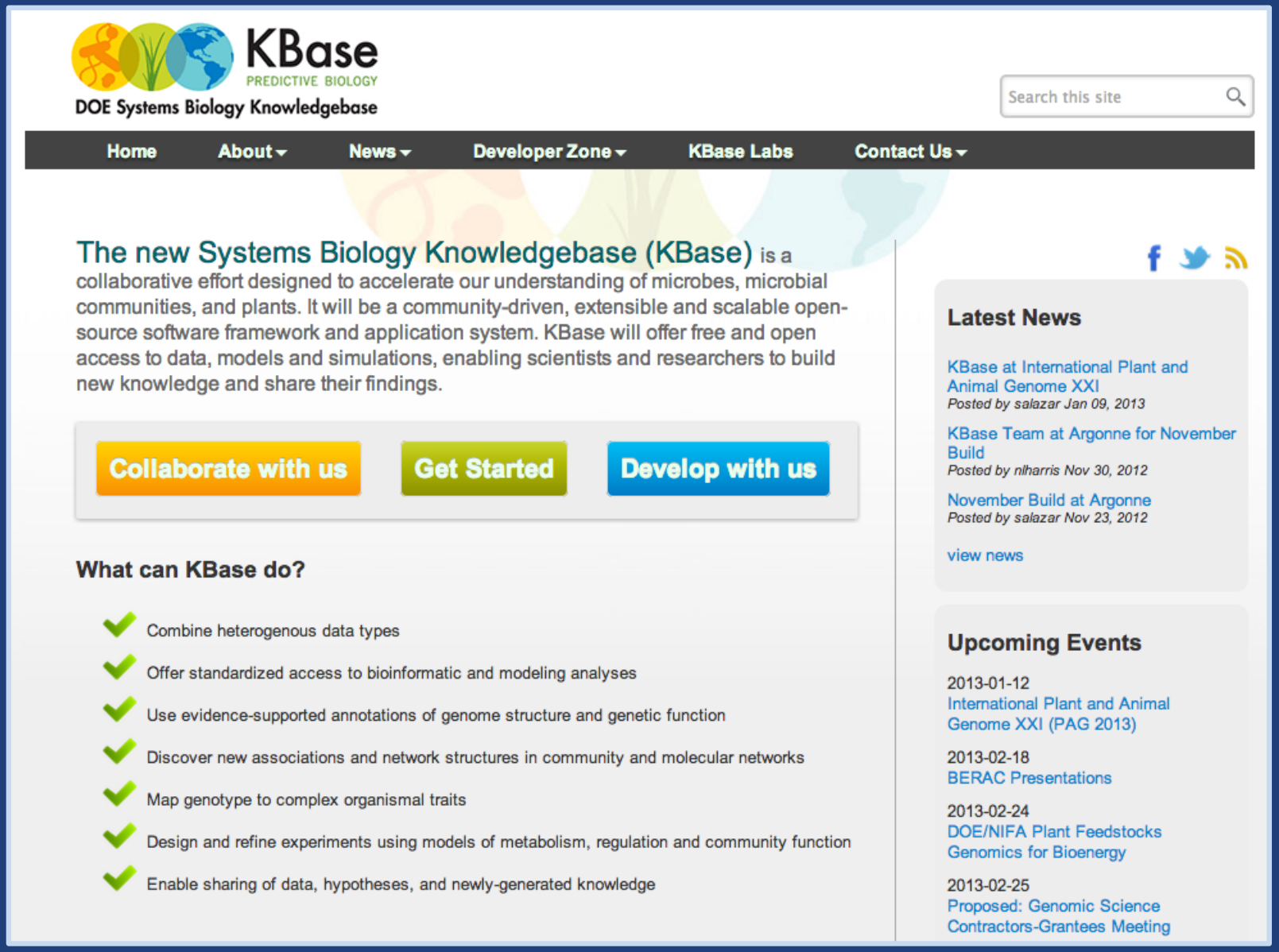
Analysis App: Wellcome Sanger



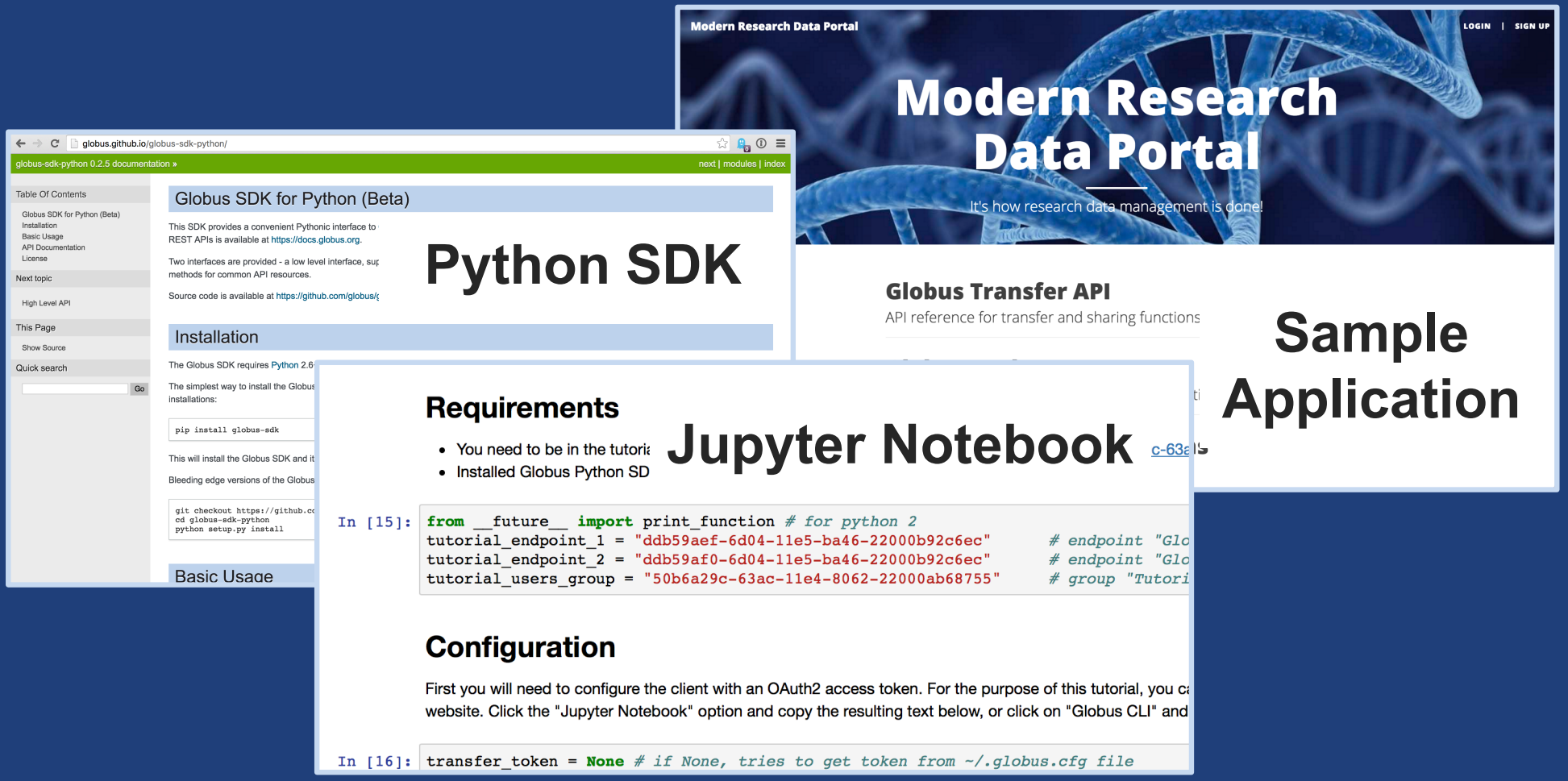
Globus PaaS: National Resource Access



Globus PaaS: Identity Management



Globus PaaS developer resources



docs.globus.org/api

github.com/globus

HTTPS support (coming soon)

- **Synchronous alternative to GridFTP**
- **Same fine-grained access control model**
- **Greatly simplified sharing/transfer of "small" datasets**
- **Standard browser behaviors**
- **Integration with clients and web apps to further leverage existing research storage systems**

Thank you to our sponsors...

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

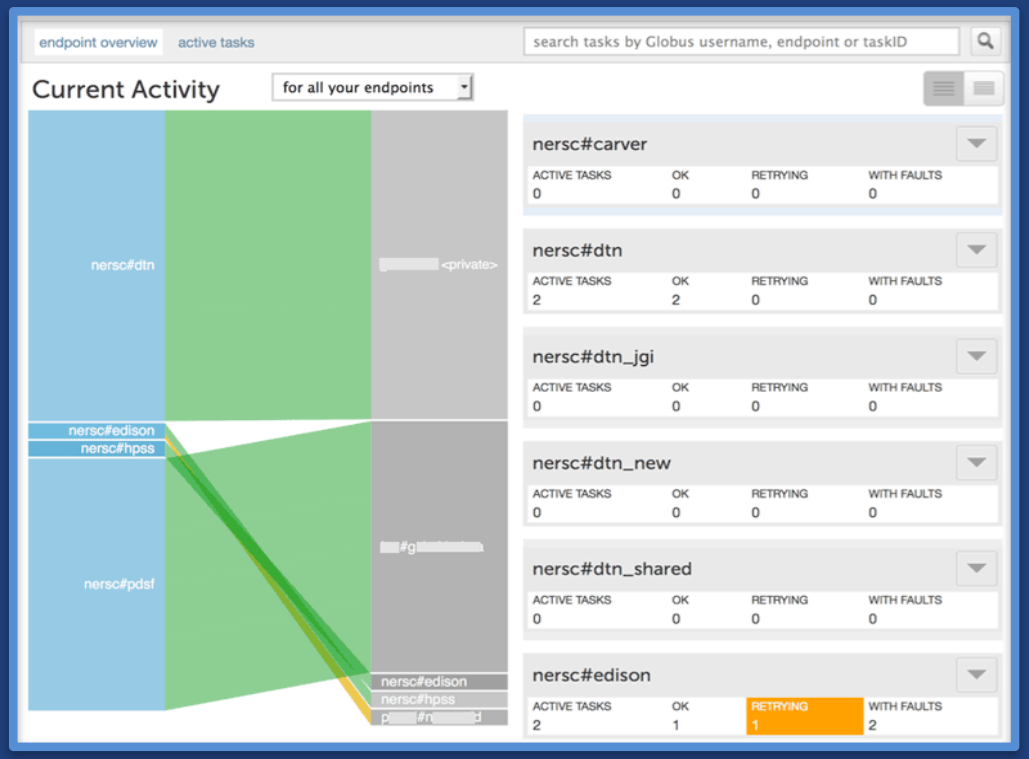
NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

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	**280PB** transferred	**47 Bn** files processed	**60,000** 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!

NERSC

NEW YORK UNIVERSITY

Berkeley
UNIVERSITY OF CALIFORNIA

JOHNS HOPKINS
UNIVERSITY

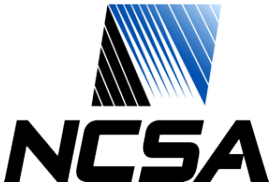
CORNELL
UNIVERSITY

UF | UNIVERSITY of FLORIDA

MICHIGAN STATE
UNIVERSITY

Yale

THE UNIVERSITY OF
CHICAGO



syngenta

NIST

Argonne
NATIONAL LABORATORY

Join the Globus community

- Access the service: **globus.org/login**
- Create a personal endpoint: **globus.org/app/endpoints/create-gcp**
- Documentation: **docs.globus.org**
- Engage: **globus.org/mailing-lists**
- Subscribe: **globus.org/subscriptions**
- Need help? **support@globus.org**
- Follow us: **@globusonline**

Globus for System Administrators

Vas Vasiliadis
vas@uchicago.edu

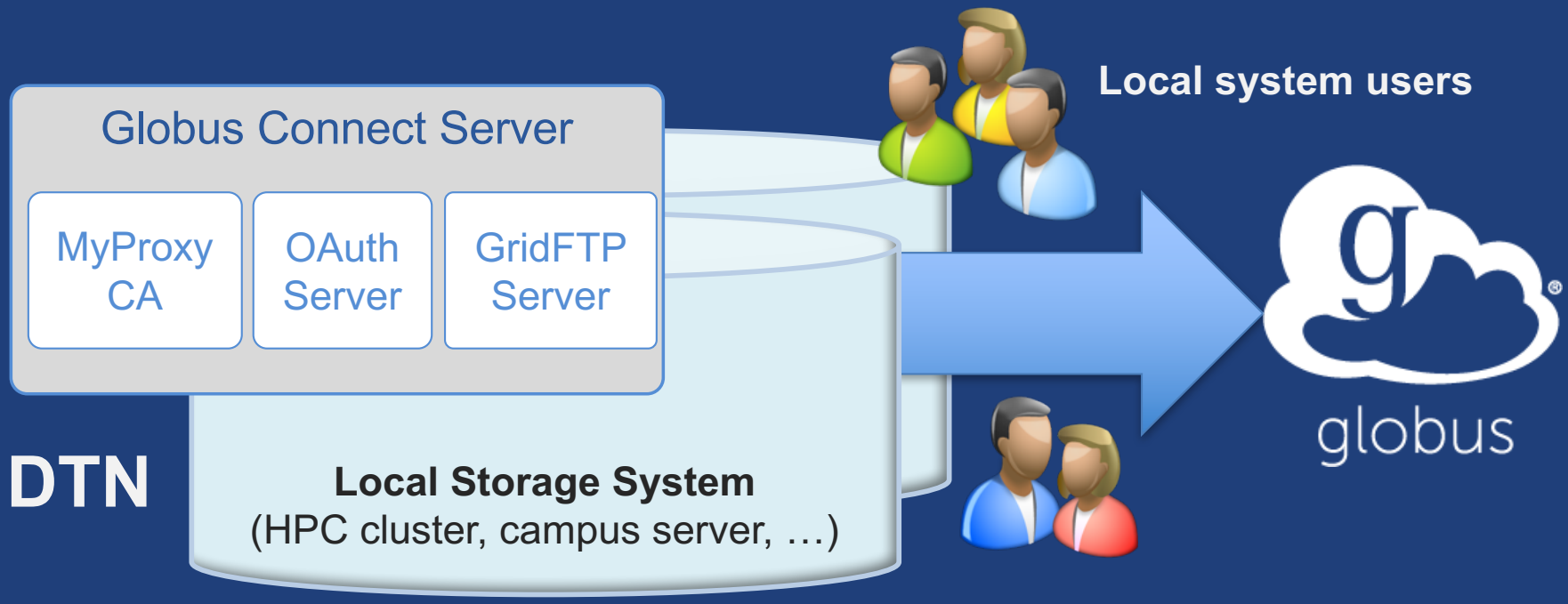
Rachana Ananthakrishnan
rachana@globus.org

Penn State University — June 29, 2017

Enabling your storage system:

Globus Connect Server

Globus Connect Server



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

Storage connectors

- **Standard storage connectors (POSIX)**
  - Linux, Windows, MacOS
  - Lustre, GPFS, OrangeFS, etc.
- **Premium storage connectors**
  - Amazon S3
  - Google Drive
  - Spectra Logic BlackPearl
  - HPSS
  - Ceph RadosGW (S3 API)
  - iRODS
  - HDFS
- **Planned: Box, Azure, *et al***

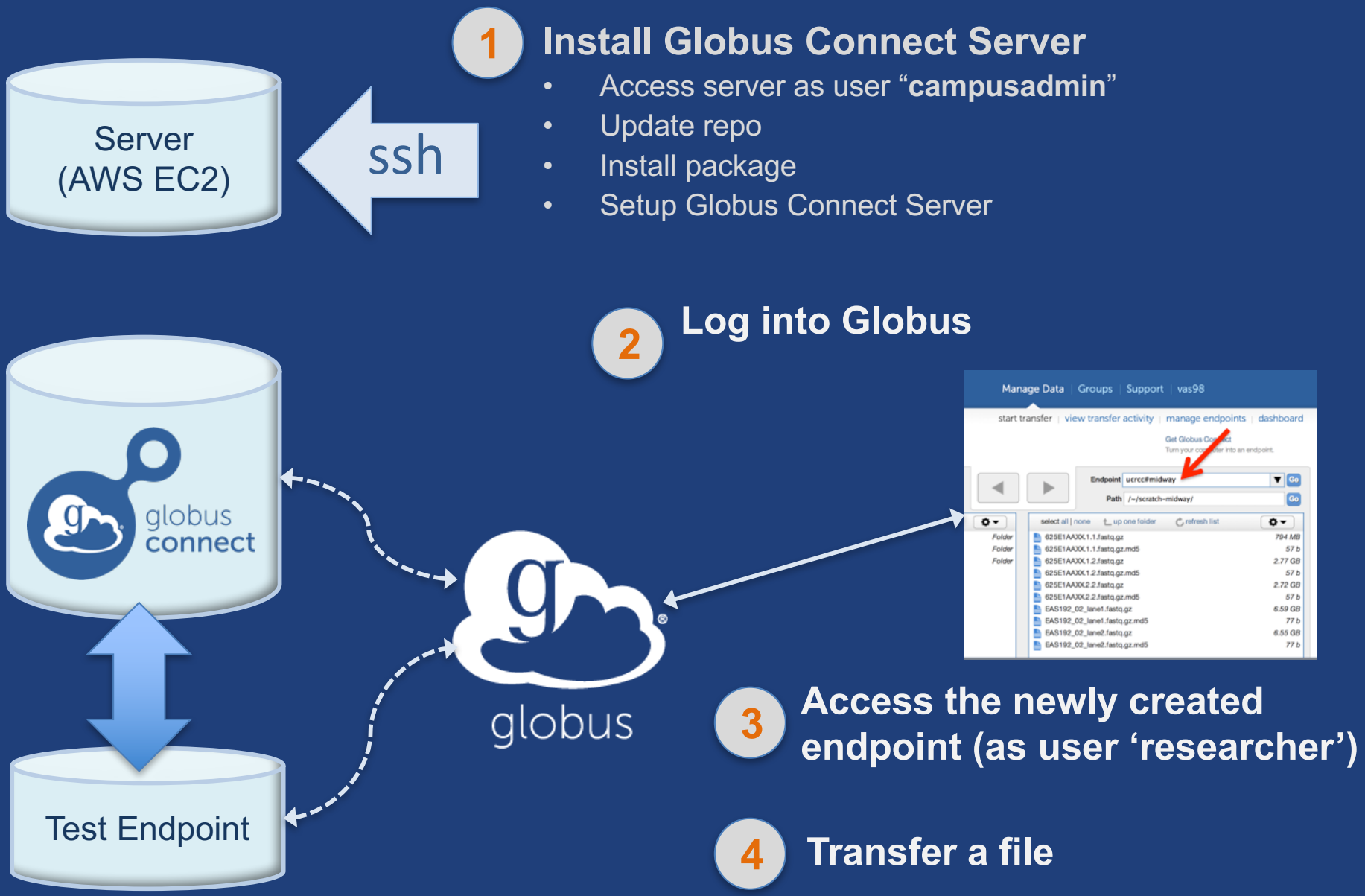
Demonstration

- **Creating a Globus endpoint on your storage system**
- **In this example, storage system = Amazon EC2 server**

Step 0: Create a Globus ID

- **Installation and configuration of Globus Connect Server requires a Globus ID**
- **Go to `globusid.org`**
- **Click “`create a Globus ID`”**

What we are going to do:



Access your host

- **Create a Globus ID**
  - Optional: associate it with your Globus account
- **Get the DNS for your EC2 server**
- **Log in as user ‘campusadmin’:**

```
ssh campusadmin@<EC2_instance_IP_address>
```

- **NB: Please** `sudo su` **before continuing**
  - User ‘campusadmin’ has sudo privileges

Step 3: Install Globus Connect Server

Cheatsheet: globusworld.org/tutorial

```
$ 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
```

Use your Globus ID username/password when prompted

You have a working Globus endpoint!

Access the Globus endpoint

- **Go to Manage Data → Transfer Files**
- **Access the endpoint you just created**
  - Search for your EC2 DNS name in the Endpoint field
  - Log in as user "**researcher**"; you should see the user's home directory
- **Transfer files to/from a test endpoint (e.g. Globus Tutorial, ESnet) and your endpoint**

Configuring Globus Connect Server

Endpoint configuration

- **Globus service config**
- **DTN (Globus Connect Server) config**
  `/etc/globus-connect-server.conf`
- **To enable changes you must run:**
  `globus-connect-server-setup`
- **“Rinse and repeat”**

Configuration file walkthrough

- **Structure based on .ini format**

```
[Section]
Option
```

- **Commonly configured options:**

```
Name
Public
RestrictedPaths
Sharing
SharingRestrictedPaths
IdentityMethod (CILogon, Oauth)
```

Exercise: Make your endpoint visible

- **Set `Public = true`**
- **Edit endpoint attributes**
  - Change the name to something useful, e.g. <your_name> EC2 Endpoint
- **Find your neighbor's endpoint**
  - You can access it too ☺

Enabling sharing on an endpoint

- **In config file, set `Sharing = True`**
- **Run `globus-connect-server-setup`**
- **Use the CLI to flag as managed endpoint**

* Note: Creation of shared endpoints requires a Globus subscription for the managed endpoint

Path Restriction

- **Default configuration:**
  - All paths allowed, access control handled by the OS
- **Use `RestrictPaths` to customize**
  - Specifies a comma separated list of full paths that clients may access
  - Each path may be prefixed by R (read) and/or W (write), or N (none) to explicitly deny access to a path
  - '~' for authenticated user's home directory, and * may be used for simple wildcard matching.
- **e.g. Full access to home directory, read access to /data:**
  - `RestrictPaths = RW~,R/data`
- **e.g. Full access to home directory, deny hidden files:**
  - `RestrictPaths = RW~,N~/.*`

Exercise: Restrict access

- **Set** `RestrictPaths=RW~,N~/archive`
- **Run** `globus-connect-server-setup`
- **Access your endpoint as** ‘researcher’
- **What’s changed?**

Limit sharing to specific accounts

- `SharingUsersAllow =`
- `SharingGroupsAllow =`
- `SharingUsersDeny =`
- `SharingGroupsDeny =`

Sharing Path Restriction

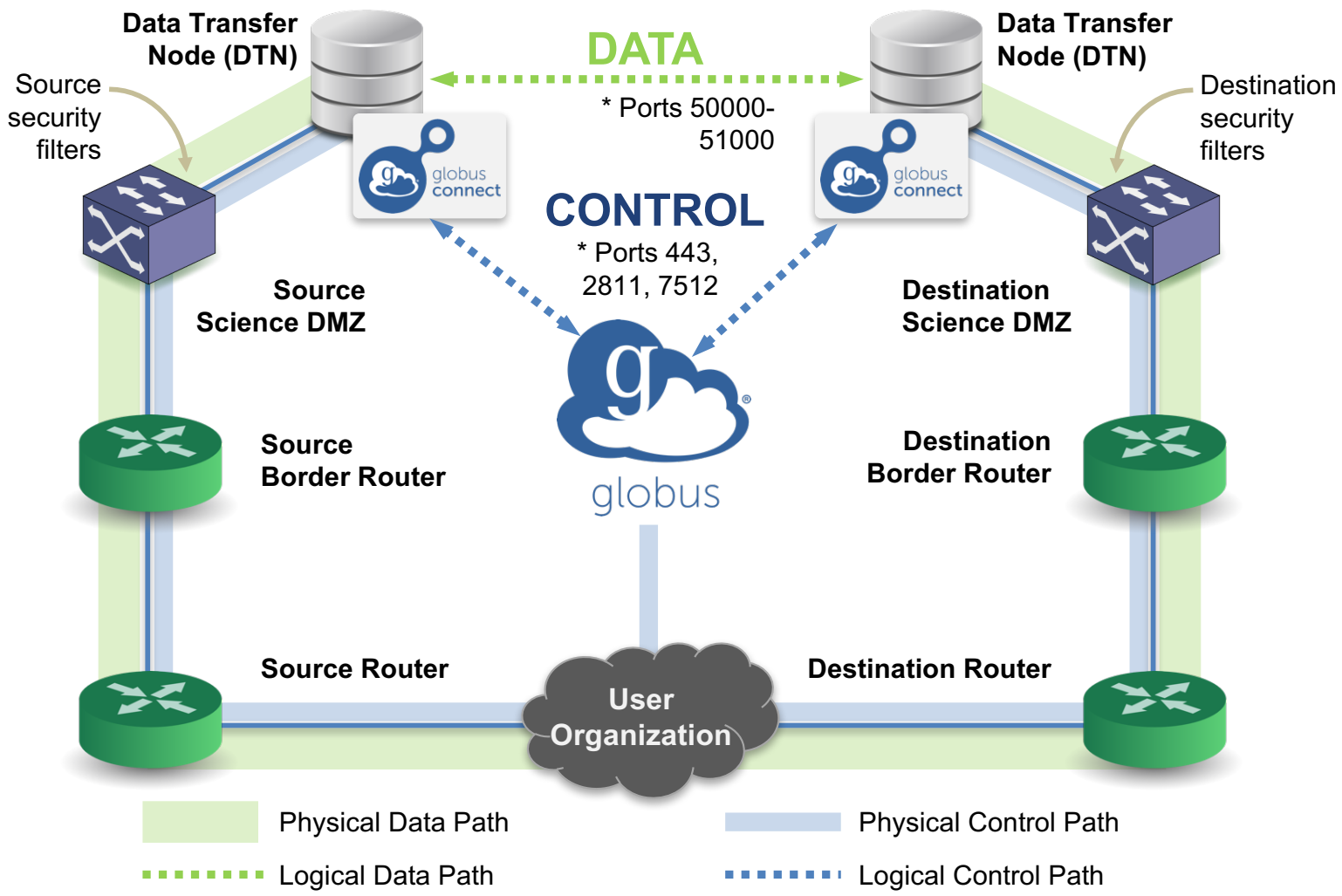
- **Restrict paths where users can create shared endpoints**
- **Use `SharingRestrictPaths` to customize**
  - Same syntax as `RestrictPaths`
- **e.g. Full access to home directory, deny hidden files:**
  - `SharingRestrictPaths = RW~,N~/.*`
- **e.g. Full access to public folder under home directory:**
  - `SharingRestrictPaths = RW~/public`
- **e.g. Full access to /proj, read access to /scratch:**
  - `SharingRestrictPaths = RW/proj,R/scratch`

Accessing Endpoints

Ports needed for Globus

- **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)

Network Paths

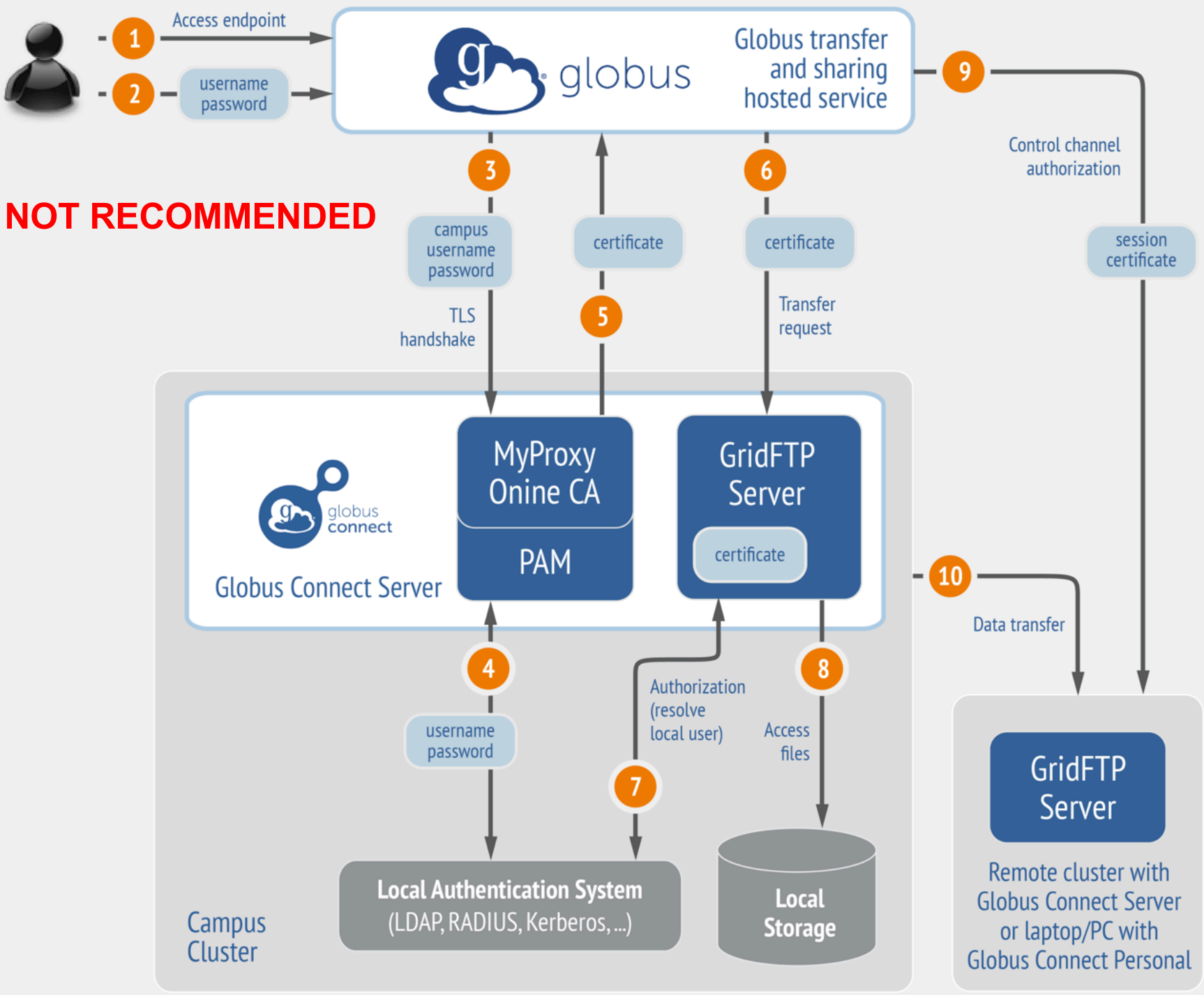


* Please see TCP ports reference: https://docs.globus.org/resource-provider-guide/#open-tcp-ports_section

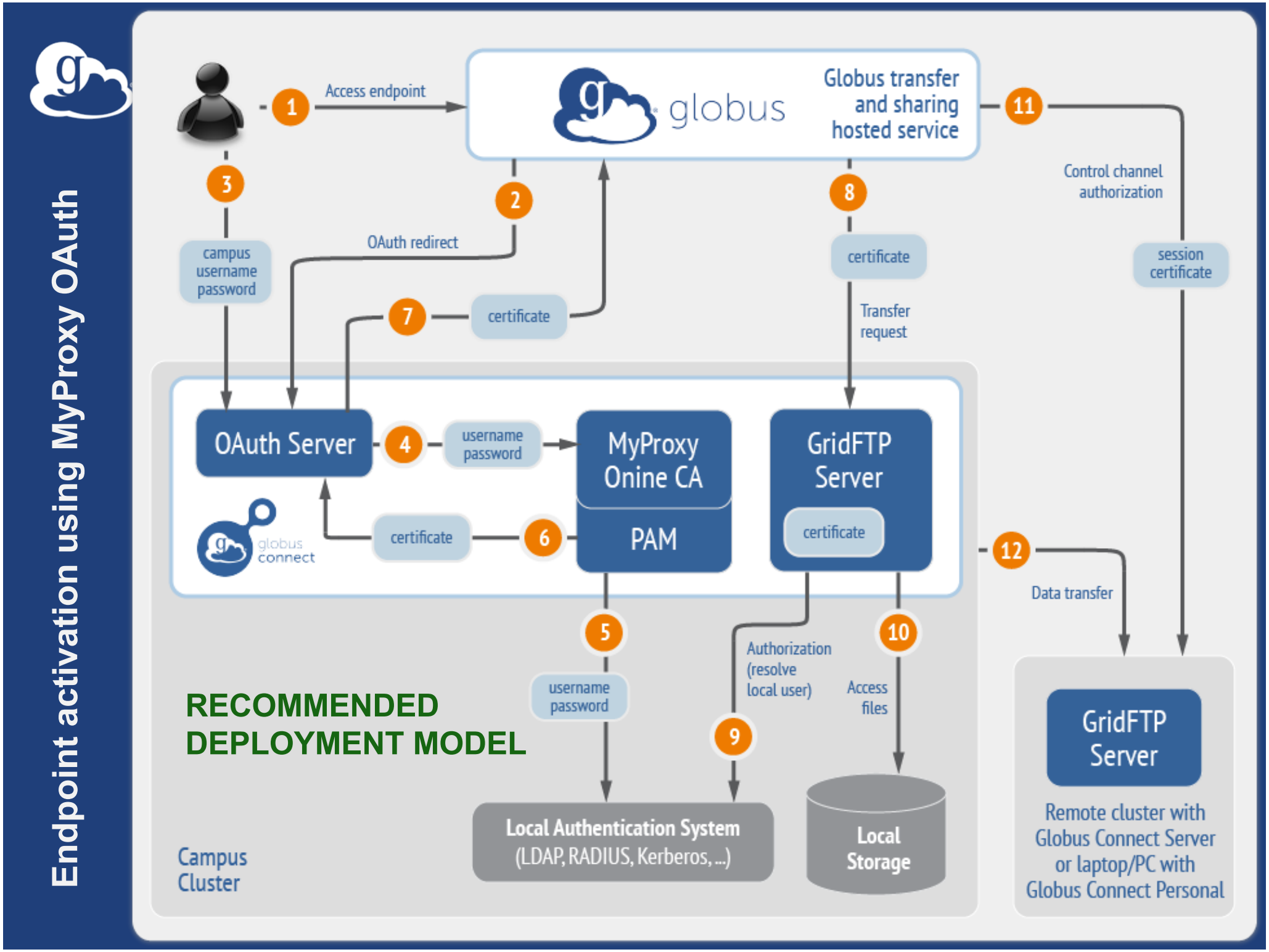
Network paths

- **Separate control and data interfaces**
- **"DataInterface =" option in globus-connect-server-conf**
- **Common scenario: route data flows over Science DMZ link**

Endpoint activation using MyProxy



Endpoint activation using MyProxy OAuth



Single Sign-On with InCommon/CILogon

- **Your Shibboleth server must release the ePPN attribute to CILogon**
- **Local resource account names must match institutional ID (InCommon ID)**
- `AuthorizationMethod = CILogon`
- `CILogonIdentityProvider = <institution_listed_in_CILogon_IdP_list>`

Integrating your IdP

- **InCommon members**
  - Must release R&S attributes to CILogon
  - Mapping uses ePPN; can use GridMap

```
AuthorizationMethod = CILogon
CILogonIdentityProvider =
<institution_name_in_CILogon_IdP_list>
```

- **Non-members**
  - IdP must support OpenID Connect
  - Requires Alternate IdP subscription
- **Using an existing MyProxy server**

Optimizing performance

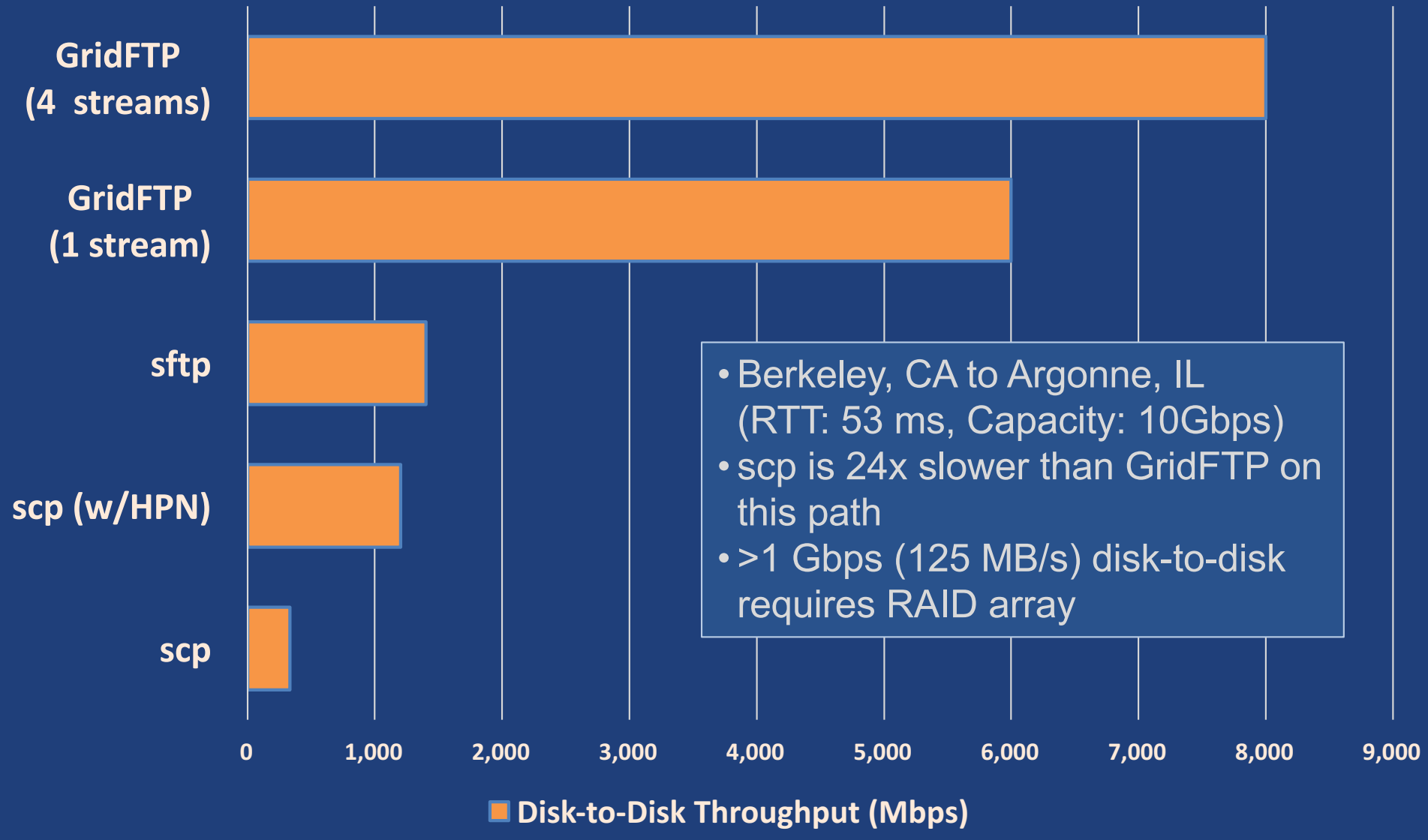
Balance: performance - reliability

- **In-flight tuning based on...**
  - Transfer request profile (#files, sizes)
  - Concurrency, parallelism
- **Request-specific overrides**
- **Endpoint-specific overrides (especially useful for multi-DTN deployments)**
- **Service limits, e.g. concurrent requests**

Illustrative performance

- **20x scp throughput (typical)**
  - >100x demonstrated
- **On par/faster than UDP based tools (NASA JPL study and anecdotal)**
- **Capable of saturating "any" WAN link**
  - Demonstrated 85Gbps sustained disk-to-disk
  - Typically require throttling for QoS

Disk-to-Disk Throughput



Source: ESnet (2016)

Network Use Parameters

- `concurrency`, `parallelism` used to configure each transfer request
- `Maximum`, `Preferred` values for each
- Use source and destination values:

```
min (max (preferred src, preferred
     dest), max src, max dest)
```

Demonstration:
Network Use Profile

Managed endpoints and subscriptions

Subscription configuration

- **Subscription manager**
  - Create/upgrade managed endpoints
  - Requires Globus ID linked to Globus account
- **Management console permissions**
  - Independent of subscription manager
  - Map managed endpoint to Globus ID
- **Globus Plus group**
  - Subscription Manager is admin
  - Can grant admin rights to other members

Creating managed endpoints

- **Required for sharing, management console, reporting, etc.**
- **Convert existing endpoint to managed:**

```
endpoint-modify --managed-endpoint <endpoint_name>
```

- **Must be run by subscription manager, using the Globus CLI**
- **Important: Re-run `endpoint-modify` after deleting/re-creating endpoint**
- **Note: The above command requires the hosted (legacy) CLI**

https://docs.globus.org/cli/#hosted_command_line_interface_legacy

Demonstration:

Command Line Interface (CLI)

Managed endpoint activity accessible via management console

- **Monitor all transfers**
- **Pause/resume specific transfers**
- **Add pause conditions with various options**
- **Resume specific tasks overriding pause conditions**
- **Cancel tasks**
- **View sharing ACLs**

Endpoint Roles

- **Administrator**: define endpoint and roles
- **Access Manager**: manage permissions
- **Activity Manager**: perform control tasks
- **Activity Monitor**: view activity

Demonstration:
Management console

Other Deployment Options

Encryption

- **Requiring encryption on an endpoint**
  - User cannot override
  - Useful for “sensitive” data
- **Globus uses OpenSSL cipher stack as currently configured on your DTN**
- **FIPS-140-2 compliance**
  - Limit number of ciphers used by OpenSSL
  - https://access.redhat.com/solutions/137833

Distributing Globus Connect Server components

- **Globus Connect Server components**
  - globus-connect-server-io, -id, -web
- **Default: -io, –id and –web on single server**
- **Common options**
  - Multiple –io servers for load balancing, failover, and performance
  - No -id server, e.g. third-party IdP such as CILogon
  - -id on separate server, e.g. non-DTN nodes
  - -web on either –id server or separate server for OAuth interface

Setting up multiple –io servers

- **Guidelines**
  - Use the same .conf file on all servers
  - First install on the server running the –id component, then all others

1. **Install Globus Connect Server on all servers**
2. **Edit .conf file on one of the servers and set `[MyProxy] Server` to the hostname of the server you want the –id component installed on**
3. **Copy the configuration file to all servers**
   - `/etc/globus-connect-server.conf`
4. **Run `globus-connect-server-setup` on the server running the –id component**
5. **Run `globus-connect-server-setup` on all other servers**
6. **Repeat steps 2-5 as necessary to update configurations**

Example: Two-node DTN

-id
-io

```
/etc/globus-connect-server.conf
[Endpoint] Name = globus_dtn
[MyProxy] Server = ec2-34-20-29-57.compute-1.amazonaws.com
```

-io

```
/etc/globus-connect-server.conf
[Endpoint] Name = globus_dtn
[MyProxy] Server = ec2-34-20-29-57.compute-1.amazonaws.com
```

For the very brave...

Globus Network Manager

- **Information from GridFTP to facilitate dynamic network changes**
- **Callbacks during GridFTP execution on local DTN**
- **Supplements information available via Globus transfer API**

Globus Network Manager Callbacks

- **Pre-listen (binding of socket)**
- **Post-listen**
- **Pre-accept/Pre-connect (no Data yet)**
- **Post-accept/Post-connect (data in flight)**
- **Pre-close**
- **Post-close**

Network manager use cases

- **Science DMZ Traffic Engineering**
  - Use SDN to dynamically route data path
  - Control path uses traditional route
- **Automated WAN bandwidth reservation**
  - OSCARS, AL2S
- **Note: All this requires custom code**

Discussion

Enable your storage system

- Everything you wanted to know: **docs.globus.org**
- Need help? **support.globus.org**
- Mailing Lists: **globus.org/mailing-lists**
- Subscribe to help us make Globus self-sustaining: **globus.org/provider-plans**
- Follow us: **@globusonline**