



# Introduction to Globus for New Users SaaS for Research Data Management

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

NC State – March 27, 2018





# 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 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 (...)

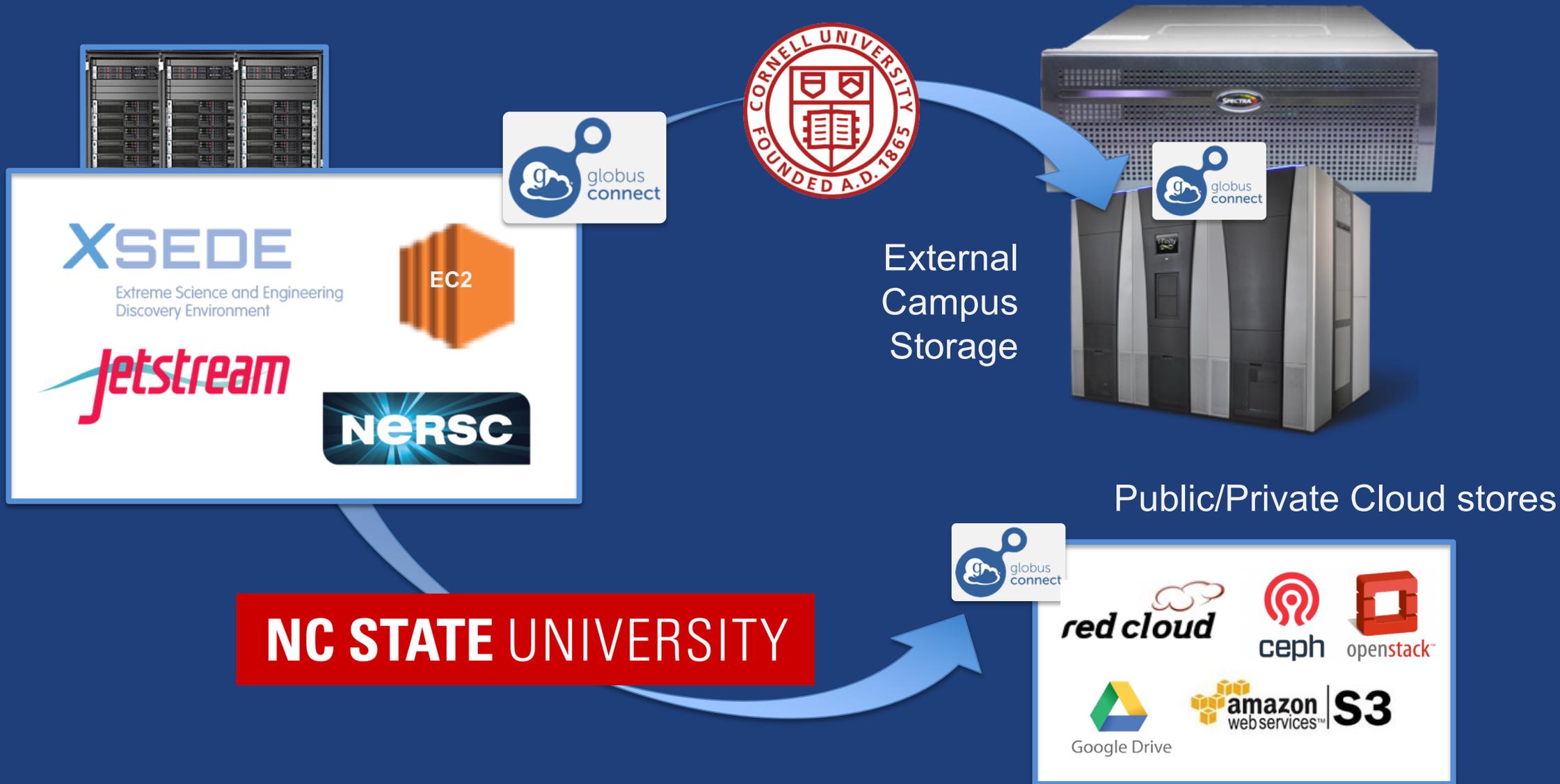


# Bridge to instruments





# Bridge to collaborators





# Bridge to community/public



Project Repositories,  
Replication Stores

**XSEDE**  
Extreme Science and Engineering  
Discovery Environment

**Jetstream**

**NERSC**

**EC2**

globus connect



Public Repositories

globus connect

globus connect

globus connect



# Globus SaaS: Research data lifecycle

Instrument



Globus transfers files reliably, securely

2

Transfer

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



Publication Repository

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

1



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

3

Share

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

6



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

5

Publish

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

8



Discover



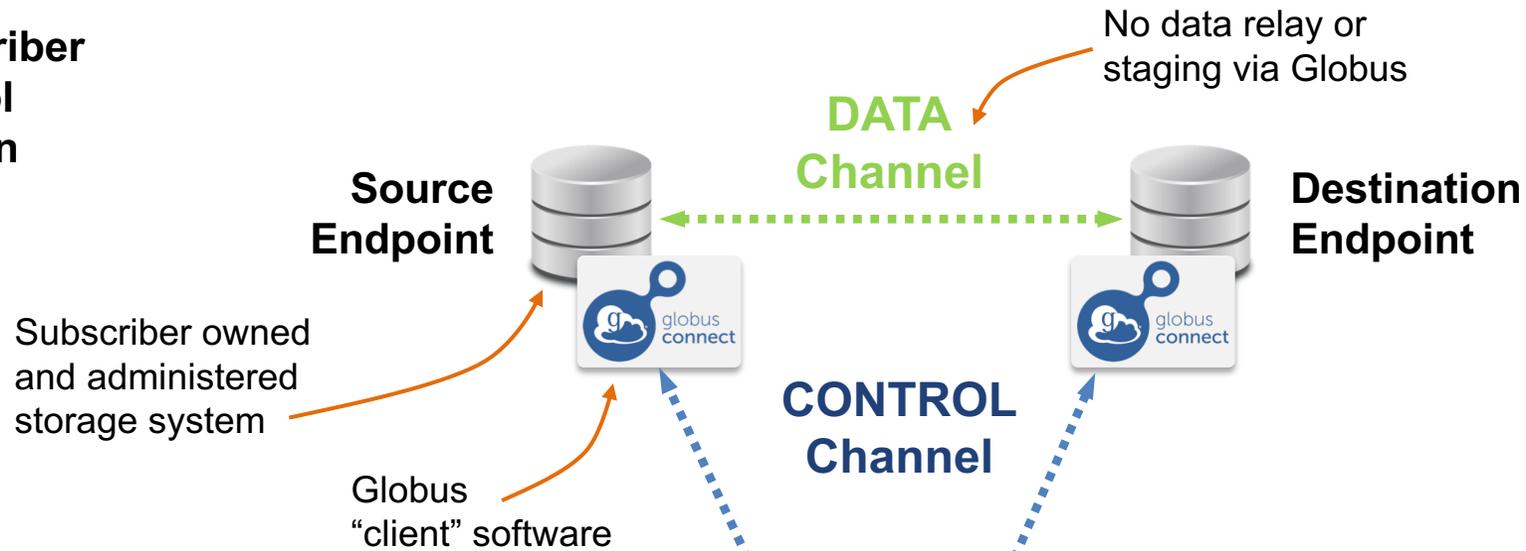
Personal Computer



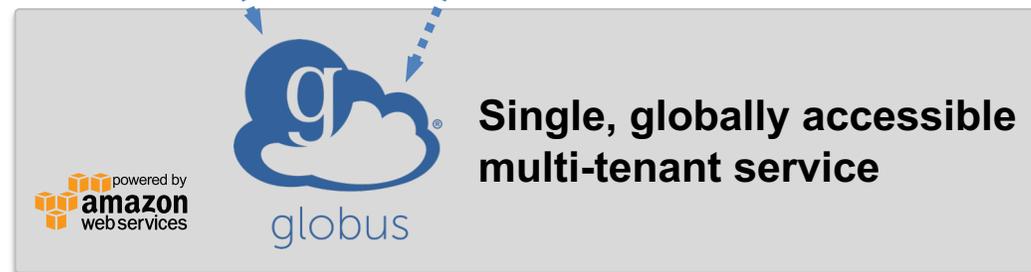
- Use a Web browser
- Access any storage
- Use an existing identity

# Conceptual architecture: Hybrid SaaS

**Subscriber  
Control  
Domain**

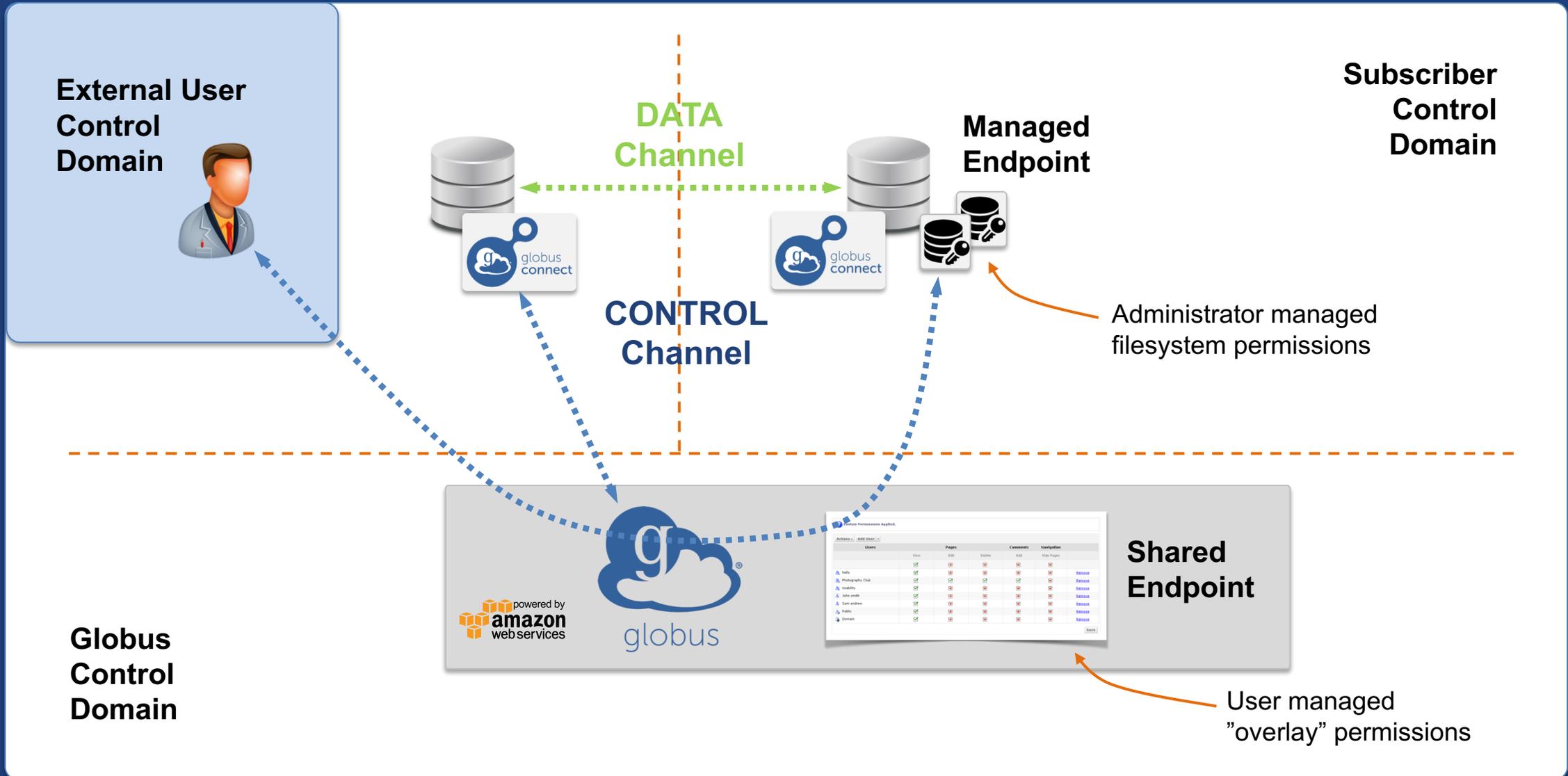


**Globus  
Control  
Domain**





# Conceptual architecture: Sharing





# 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**  
**Group Management**



# Data Publication and Discovery

The screenshot shows the web interface for the Materials Data Facility (MDF) on the Globus platform. At the top left is the Globus logo and the word "globus". To the right are "Log In" and "Sign Up" links. A light blue banner below the header contains the text: "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 heading "Materials Data Facility Community home page" followed by a large, colorful logo for "MATERIALS DATA FACILITY" composed of many small circles in shades of blue, green, yellow, and orange. Below the logo 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." This is followed by another paragraph: "MDF is a pilot project funded by NIST, and serves as the first pilot community of the [National Data Service](#)." Below that is a line of text: "Contact Ben Blaiszik ([blaiszik@uchicago.edu](mailto:blaiszik@uchicago.edu)) to begin publishing your data". At the bottom of the main content area is a "Browse" section with four buttons labeled "Issue Date", "Author", "Title", and "Subject".

<https://publish.globus.org>

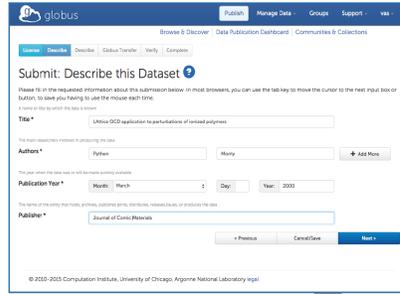


# Globus data publication framework

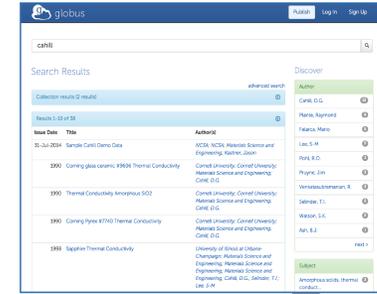
Identifier	URL	Handle	DOI	
Description	None	Standard	Domain-specific	Custom
Curation	None	Acceptance	Human-validated	Machine-validated
Access	Anonymous	Public	Embargoed	Collaborators
Preservation	Transient	Project Lifetime	Archive	“forever”



# Publish

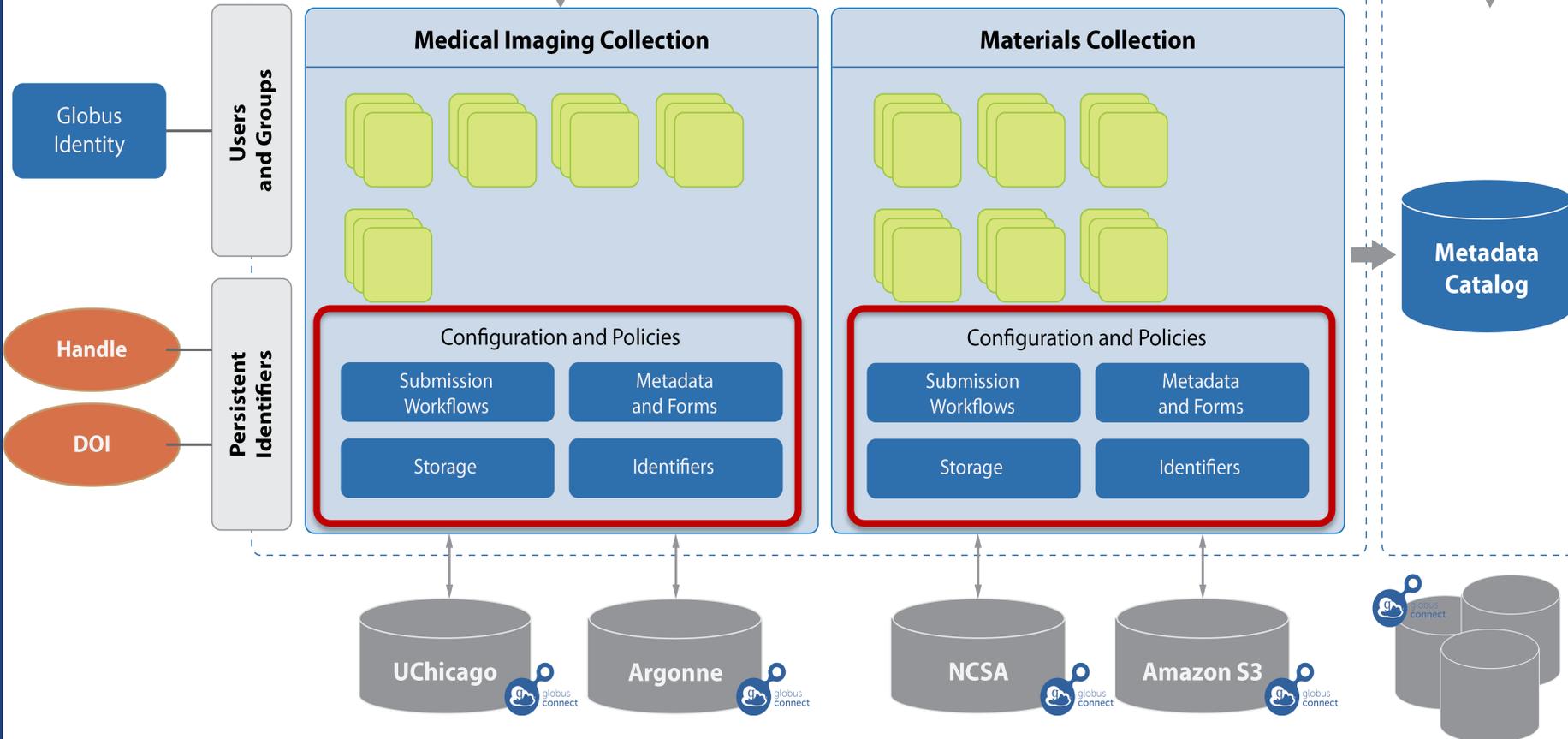


# Discover



## Globus Authentication

### Globus Data Publication





# Demonstration Data Publication



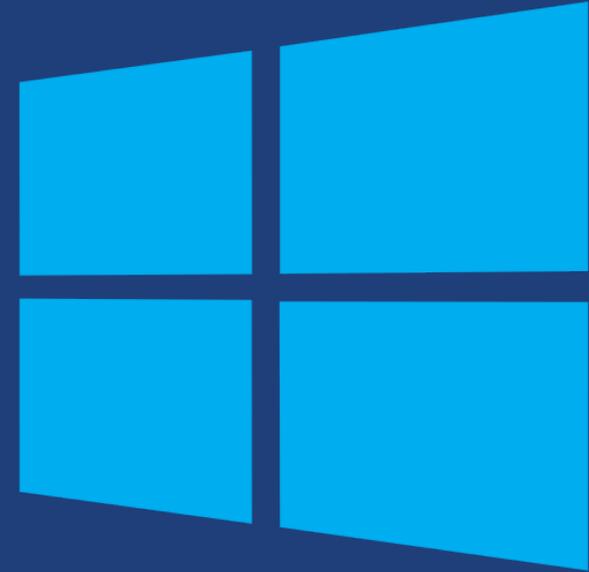
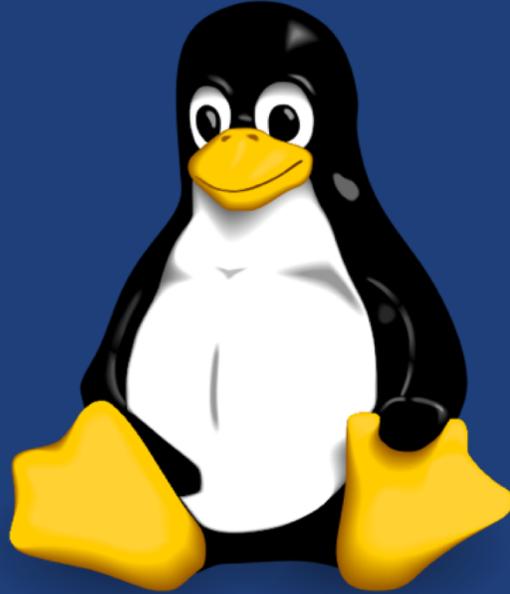
**How can I use Globus  
on my computer?**



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



# Globus Connect Personal



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



**How can I integrate  
Globus into my  
research workflows?**



**Globus serves as...**

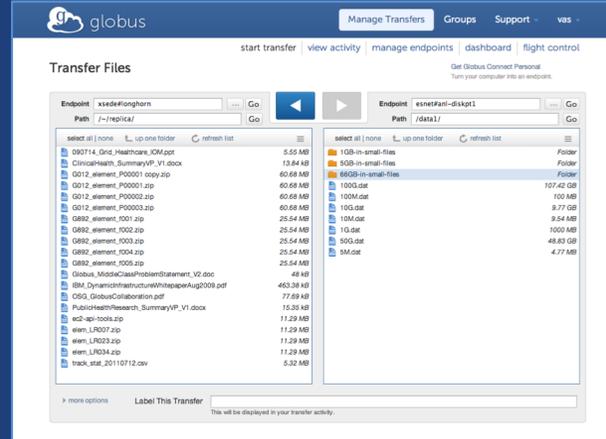
**...a platform for building  
science gateways, portals,  
and other web applications  
in support of research and  
education**



# 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.
```

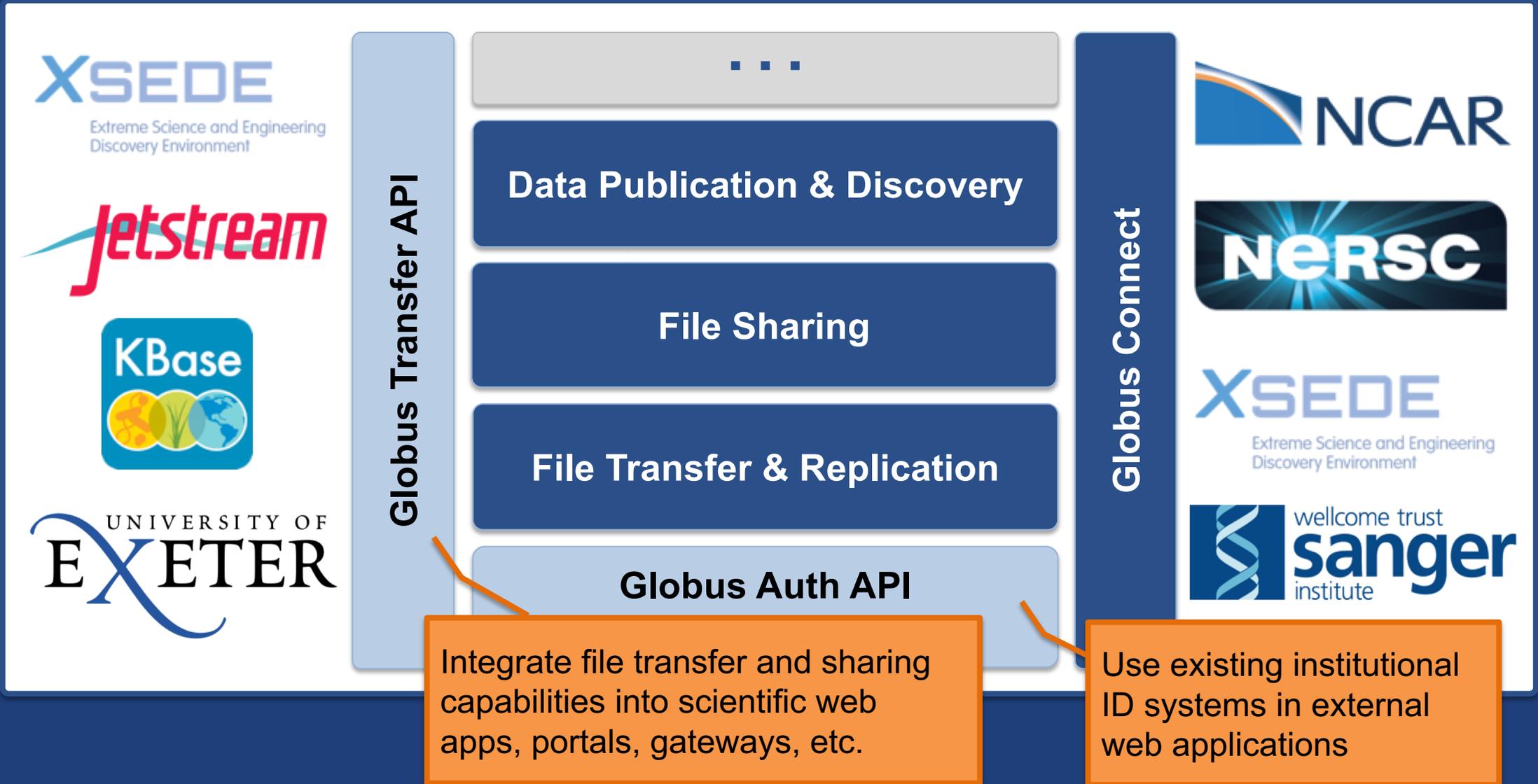
CLI

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

Rest API



# Globus as PaaS





# Globus PaaS developer resources

[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

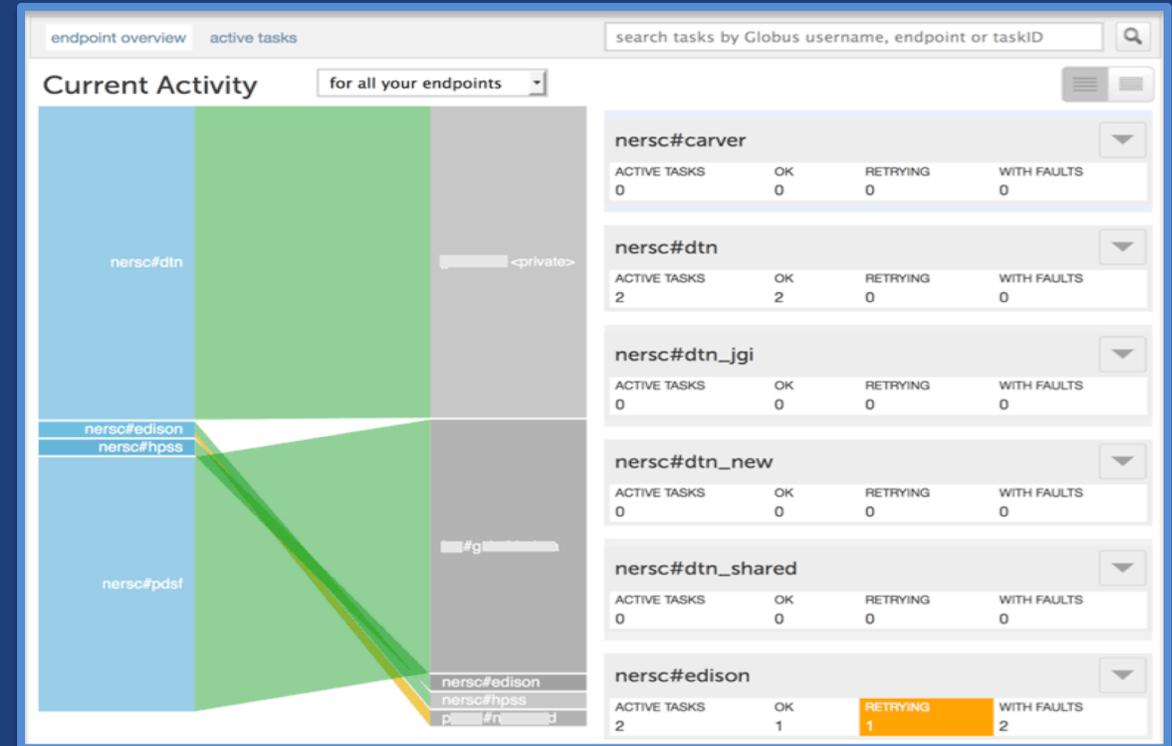
Argonne  
NATIONAL LABORATORY

powered by  
**amazon**  
web services



# Globus sustainability model

- **Standard Subscription**
  - Shared endpoints
  - Data publication
  - Management console
  - Usage reporting
  - Priority support
  - Application integration
  - HTTPS support (coming soon)
- **Branded Web Site**
- **Premium Storage Connectors**
- **Alternate Identity Provider (InCommon is standard)**





# Thank you to our users...

**48**

most server endpoints at a single organization

**384 PB**  
transferred

**64 billion**  
tasks processed

**76,000**  
registered users

**500**

100TB+ users

**14,000**  
active users

**3 months**

longest running managed transfer

**10,000**

active endpoints

**350+**

federated identities

**1 PB**

largest single transfer to date

**5,000**

active shared endpoints

**99.5%**

uptime



# Our supporters



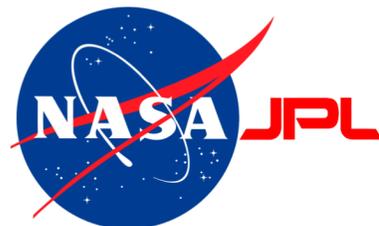
JOHNS HOPKINS  
UNIVERSITY



Yale



CORNELL  
UNIVERSITY



THE UNIVERSITY OF  
CHICAGO



MICHIGAN STATE  
UNIVERSITY



Stanford  
University



NIST

Dartmouth

SIMONS FOUNDATION





# 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)



# Help us get the word out!

- **Share your experiences!**
  - **Contribute** to our Usage Brief Library
  - **Add a slide** or logo in event talks (we can help!)
  - **Mention Globus** in news articles or interviews
  - **Tag us** in posts about projects that use Globus
  - **Acknowledge Globus** in your journal articles
- **Why?**
  - Give your peers new ideas on how to use Globus
  - Help us grow the user community

“..., and file sharing with Globus.”

“...with Globus for data transfer.”

“We used Globus for...”

“...and Globus.”

“I needed Globus to...”

“#ALCF #ORNL #theNCI #CANDLE #globusonline”

“...using tool x, tool y, Globus, technology z...”



# Managing Globus Endpoints

## Globus for System Administrators

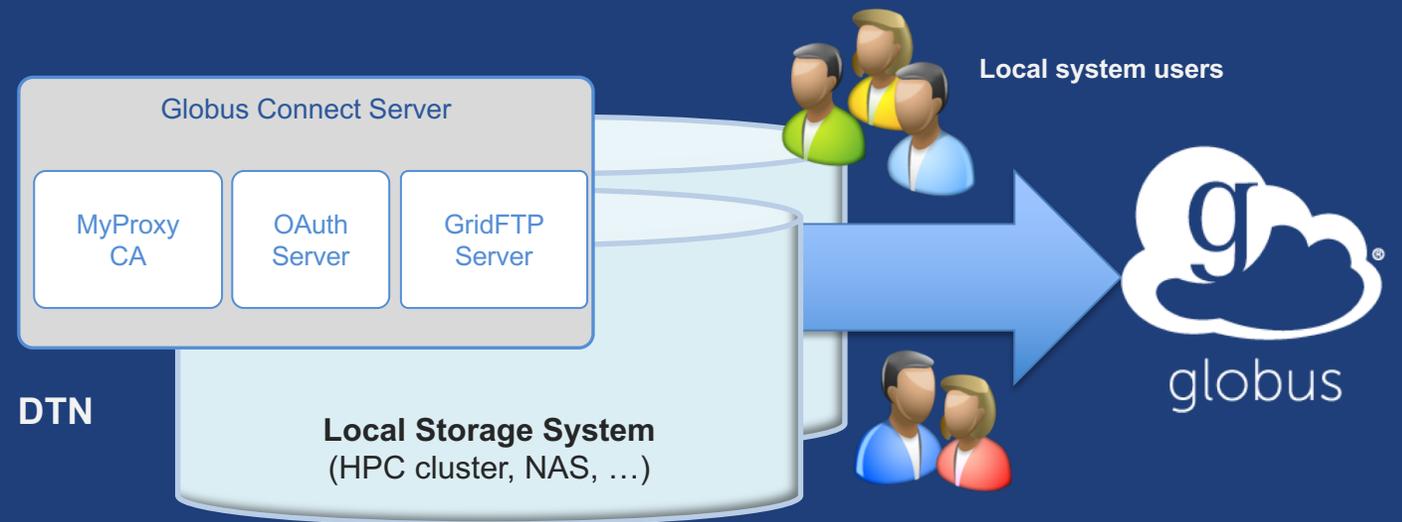
Vas Vasiliadis  
vas@uchicago.edu

NC State – March 27, 2018



# Globus Connect Server

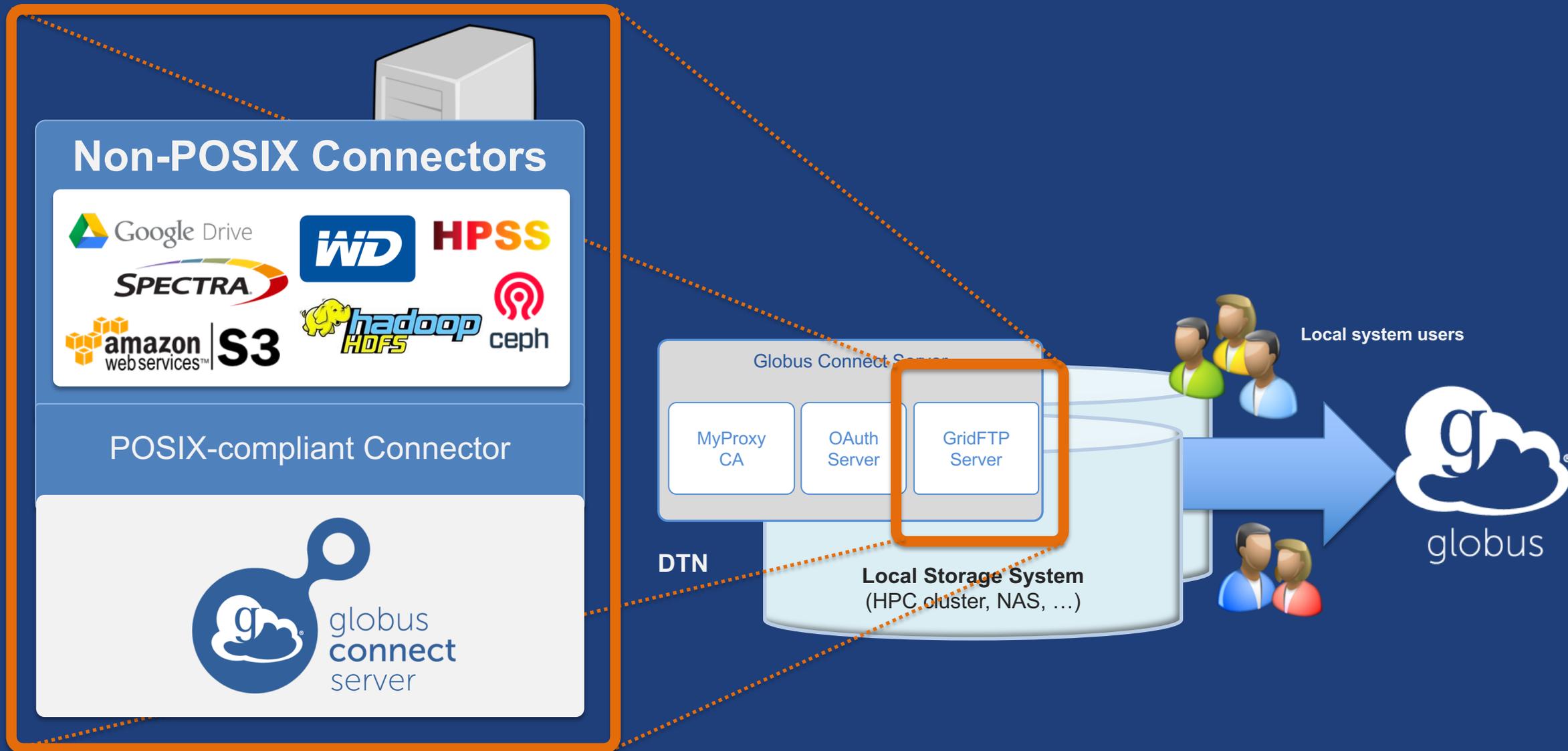
- **Makes your storage accessible via Globus**
- **Multi-user server, installed and managed by sysadmin**
- **Default access for all local accounts**
- **Native packaging  
Linux: DEB, RPM**



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



# Globus Connect Server

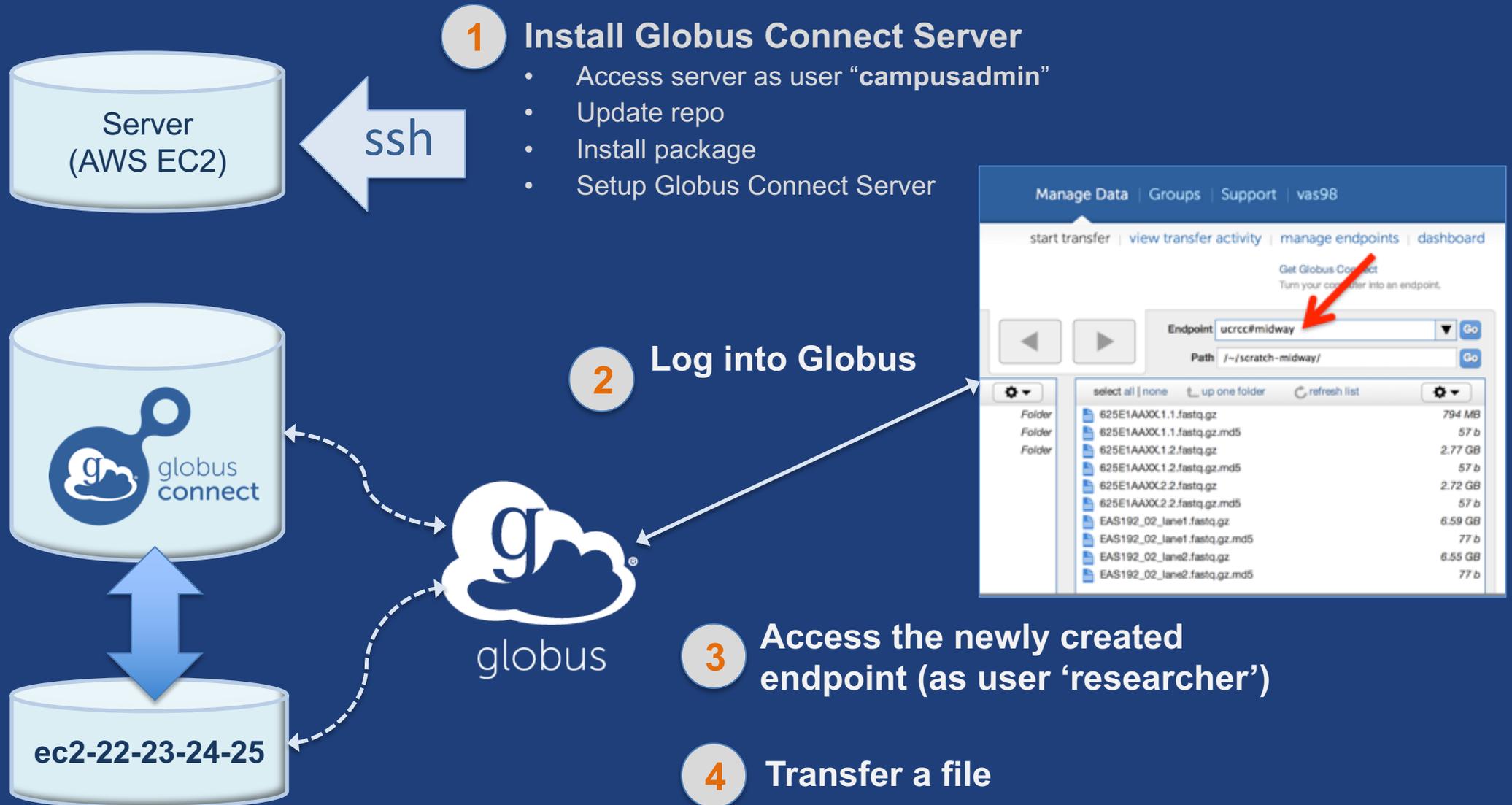


# Creating a Globus endpoint on your server

- **In this example, Server = Amazon EC2 instance**
- **Installation and configuration of Globus Connect Server requires a Globus ID**
- **Go to [globusid.org](https://globusid.org)**
- **Click “create a Globus ID”**
  - Optional: associate it with your Globus account



# What we are going to do:





# Access your host

- **Get the IP address 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



# Install Globus Connect Server

```
$ sudo su
$ curl -L0s 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 host name in the Endpoint field
  - Log in as “researcher”; you will see the user’s home directory
- **Transfer files to/from a test endpoint (e.g. Globus Tutorial) and your EC2 endpoint**



# Configuring Globus Connect Server

# Endpoint configuration

- **Globus service “Manage Endpoints” page**
- **DTN (Globus Connect Server) config**
  - `/etc/globus-connect-server.conf`
  - Standard .ini format: `[Section] Option = value`
  - To enable changes you must run:  
`globus-connect-server-setup`
  - “Rinse and repeat”



# Common configuration options

- **Manage Endpoints page**
  - Display Name
  - Visibility
  - Encryption
- **DTN configuration file – common options:**
  - RestrictPaths
  - IdentityMethod (CILogon, OAuth)
  - Sharing
  - SharingRestrictPaths



# 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~/.\*

# Enabling sharing on an endpoint

- In config file, set `Sharing=True`
- Run `globus-connect-server-setup`
- Use the CLI to flag as managed endpoint (also configurable via the web app)

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

# 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`



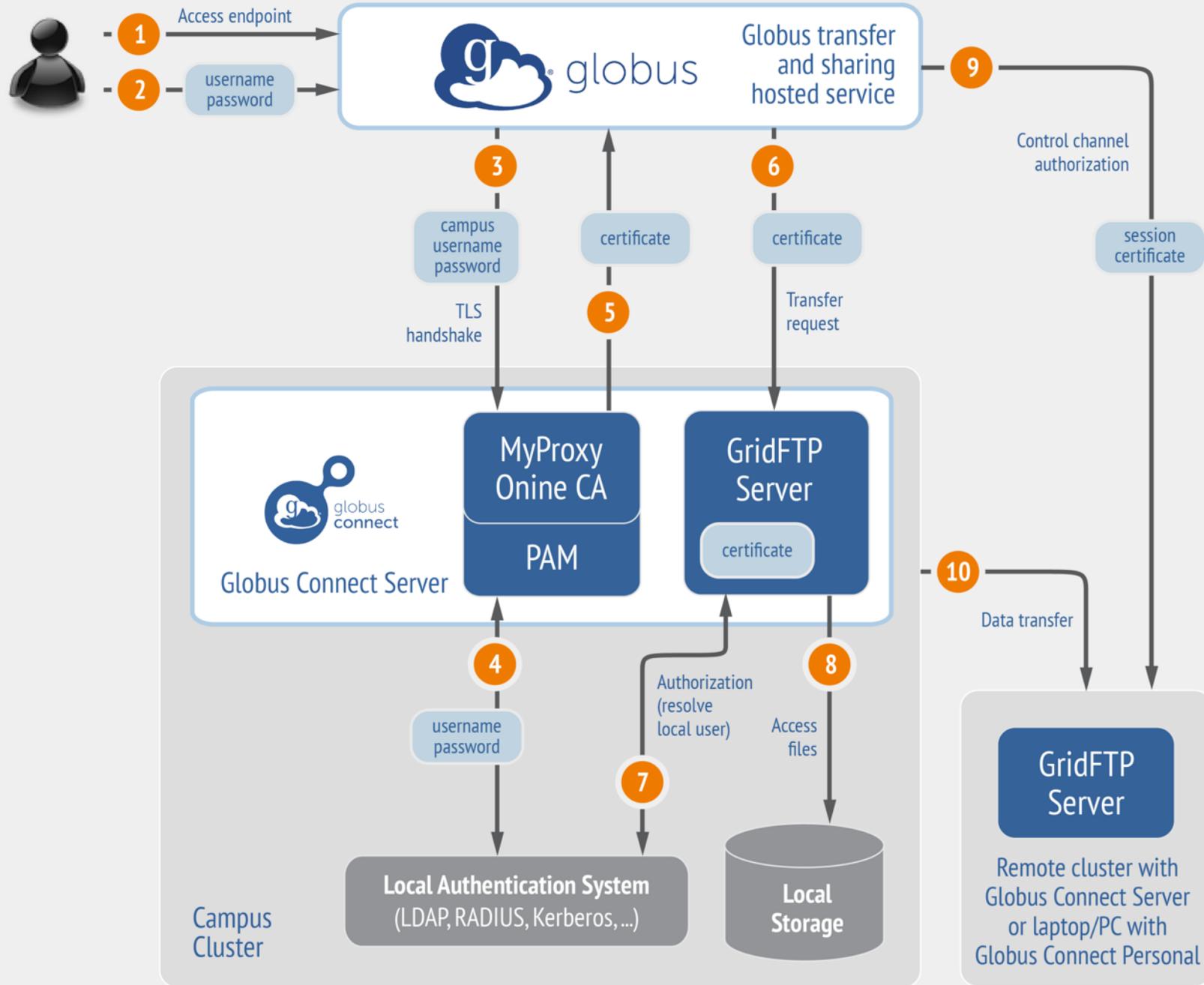
# Accessing Endpoints

# 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)



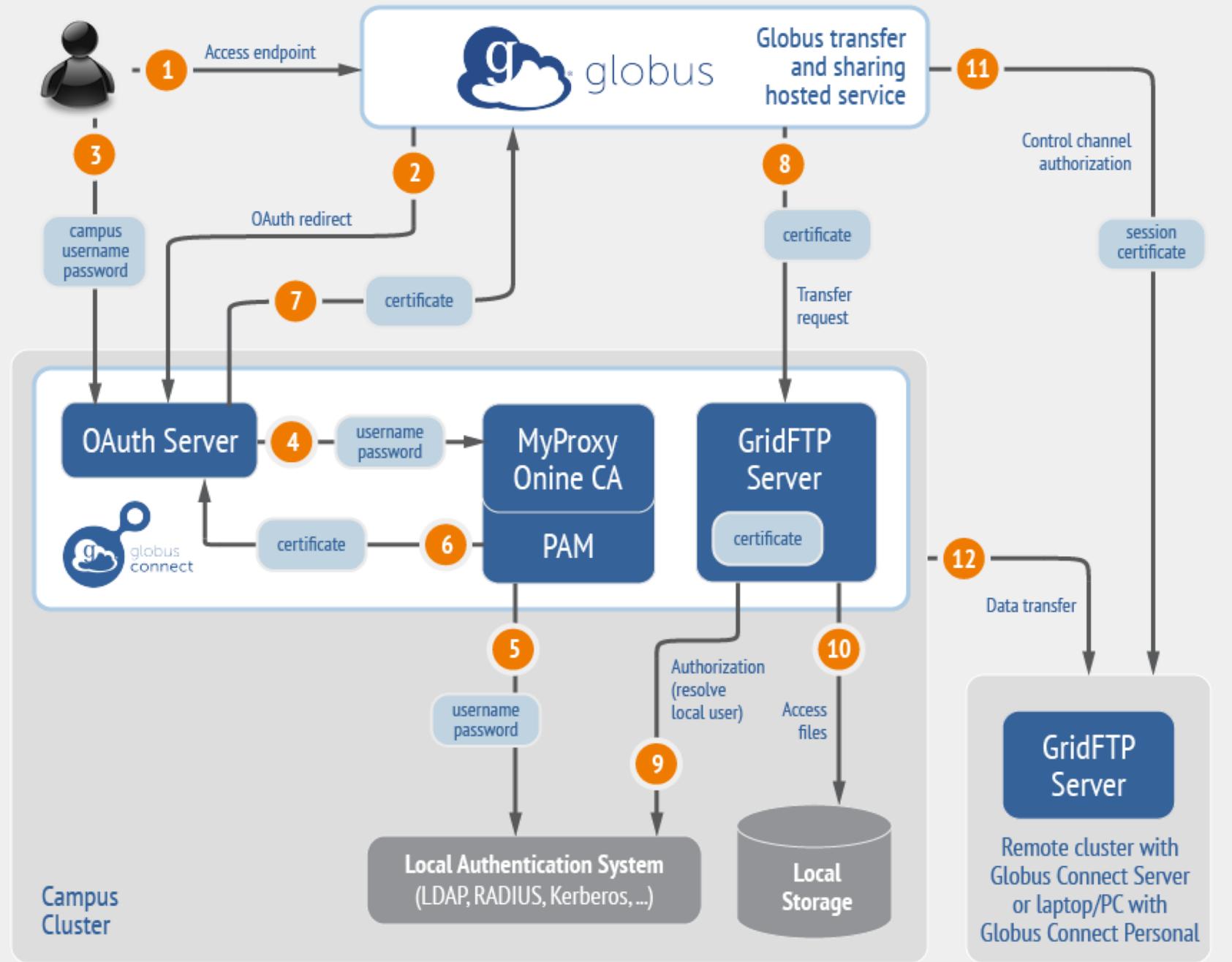
# Endpoint activation using MyProxy



Default configuration  
(avoid if at all possible)



# Endpoint activation using MyProxy OAuth



## Best practice configuration

# Single Sign-On with InCommon/CILogon

- **Your Shibboleth server must release R&S attributes to CILogon—especially the ePPN attribute**
- **Local resource account names must match your institutional ID (InCommon ID)**
- **In `/etc/globus-connect-server.conf` set:**

```
AuthorizationMethod = CILogon
```

```
CILogonIdentityProvider =  
<institution_listed_in_CILogon_IdP_list>
```



# 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, ...**
- **Convert existing endpoint to managed via CLI (or web):**  
`globus endpoint update --managed <endpt_uuid>`
- **Must be run by subscription manager**
- **Important: Re-run endpoint update after deleting/re-creating endpoint**



# Monitoring and managing Globus endpoint activity

# 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**  
**Endpoint Roles**  
**Usage Reporting**



**...on performance**

# Balance: performance - reliability

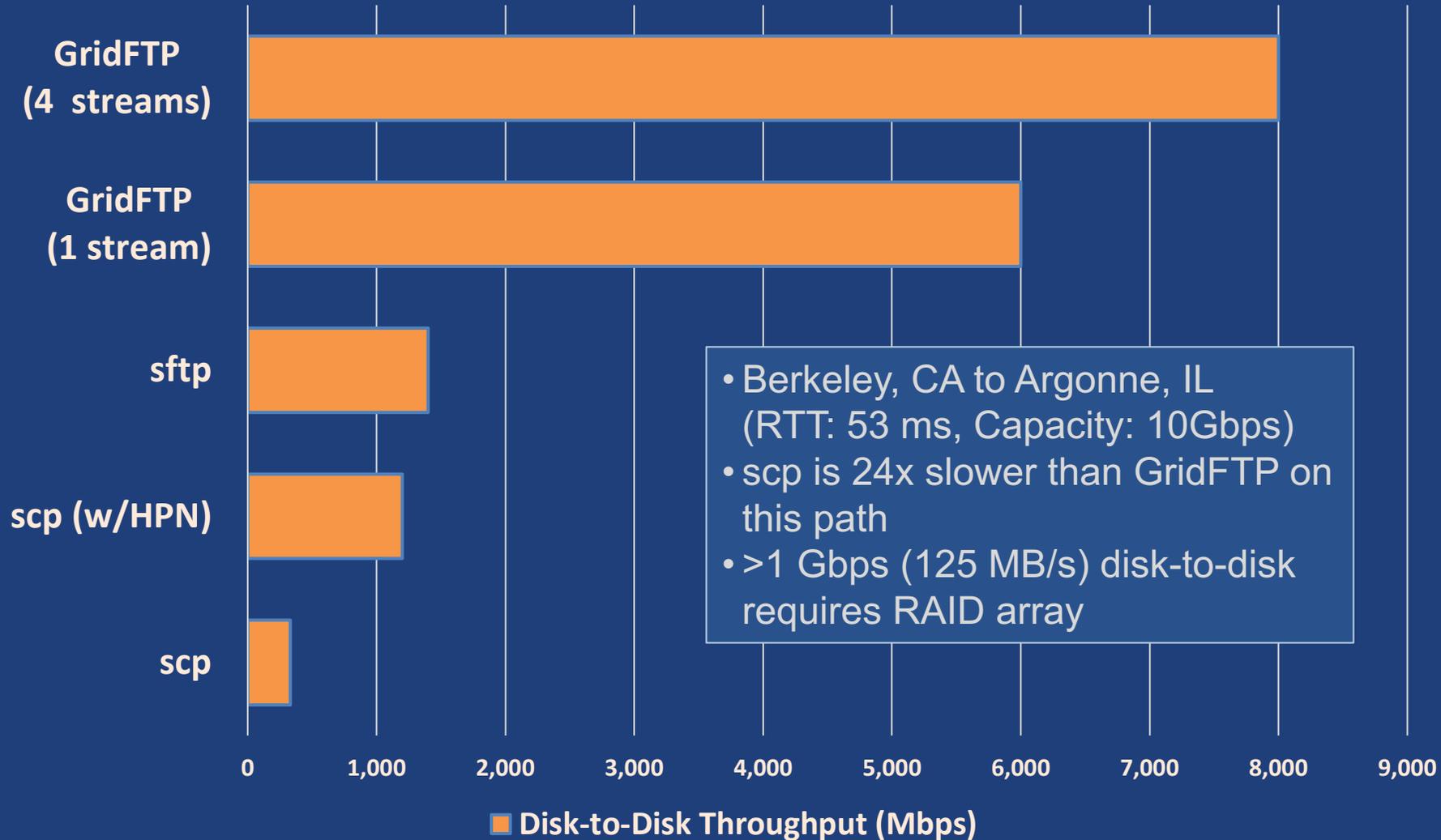
- **Network use parameters: concurrency, parallelism**
- **Maximum, Preferred values for each**
- **Transfer considers source and destination endpoint settings**

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

- **Service limits, e.g. concurrent requests**



# Disk-to-Disk Throughput: ESnet Testing

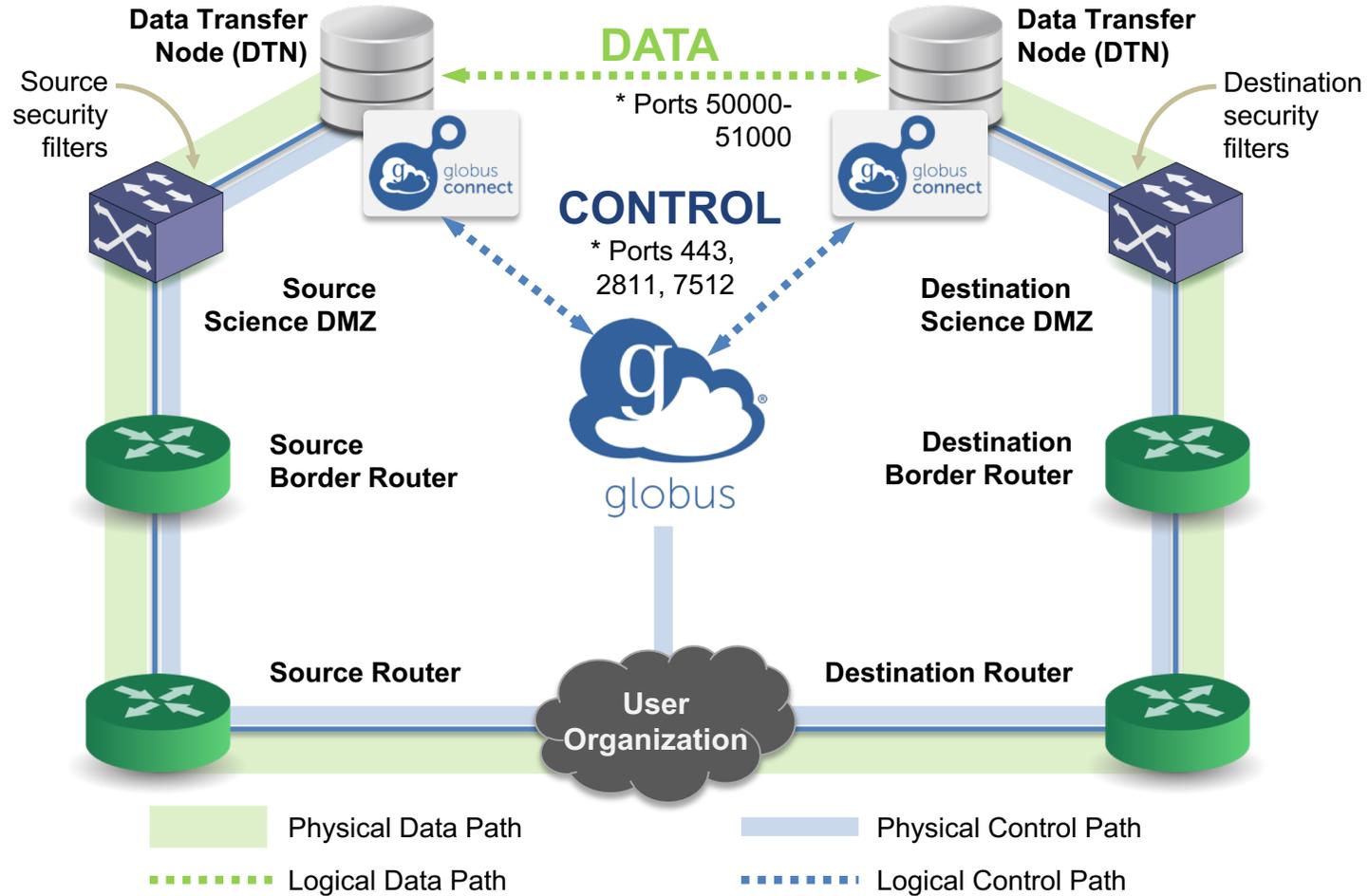




# Deployment Scenarios



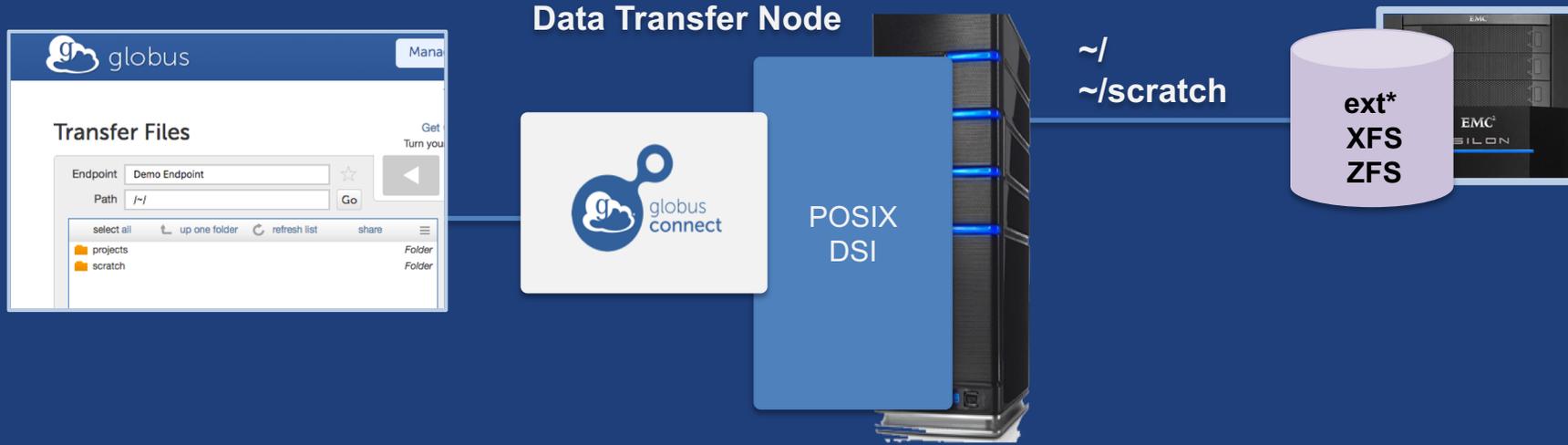
# Best practice network configuration



\* 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)

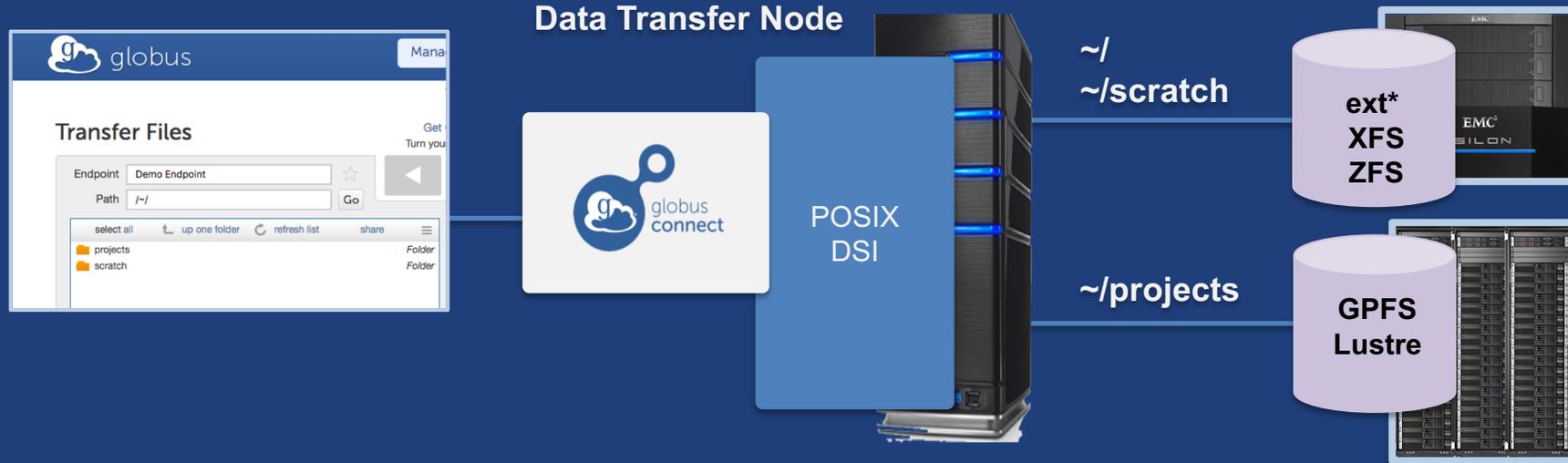


# Multi-endpoint configuration



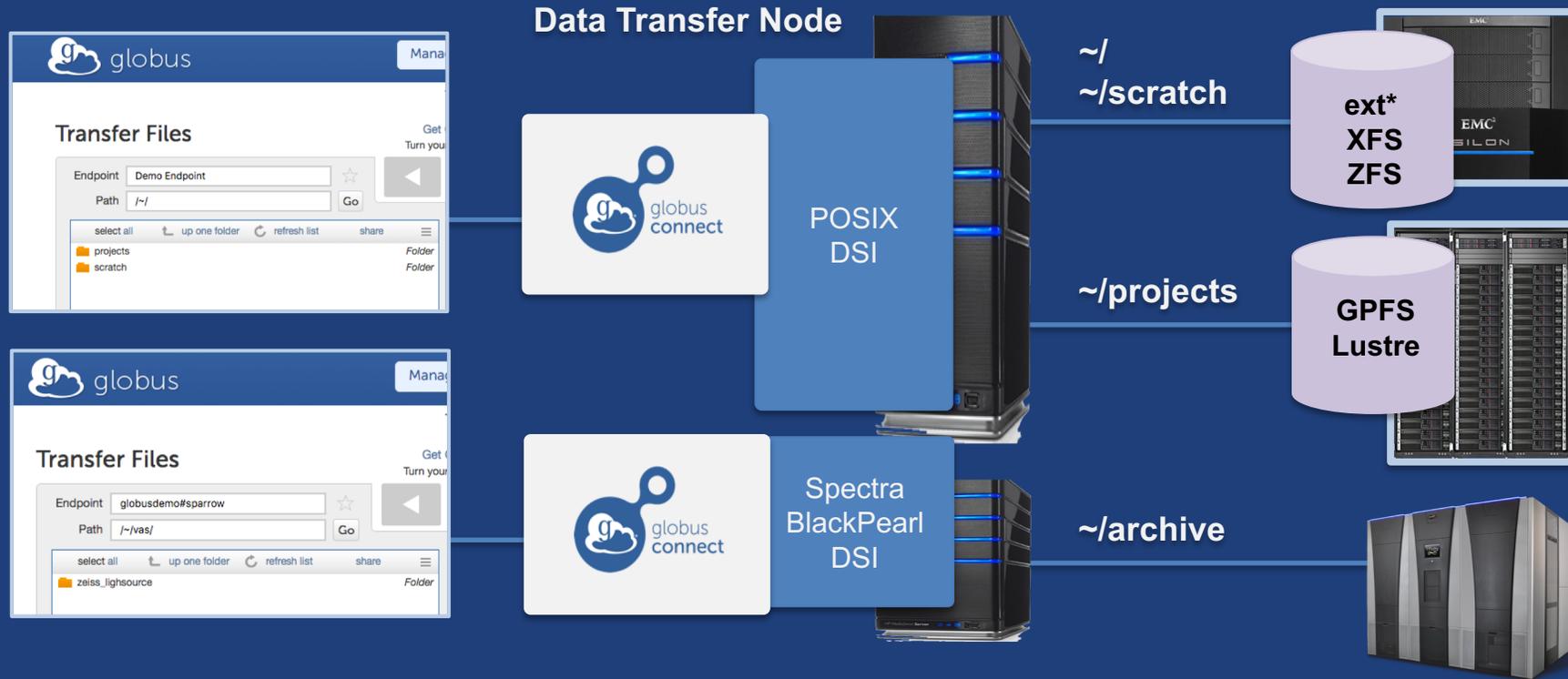


# Multi-endpoint configuration



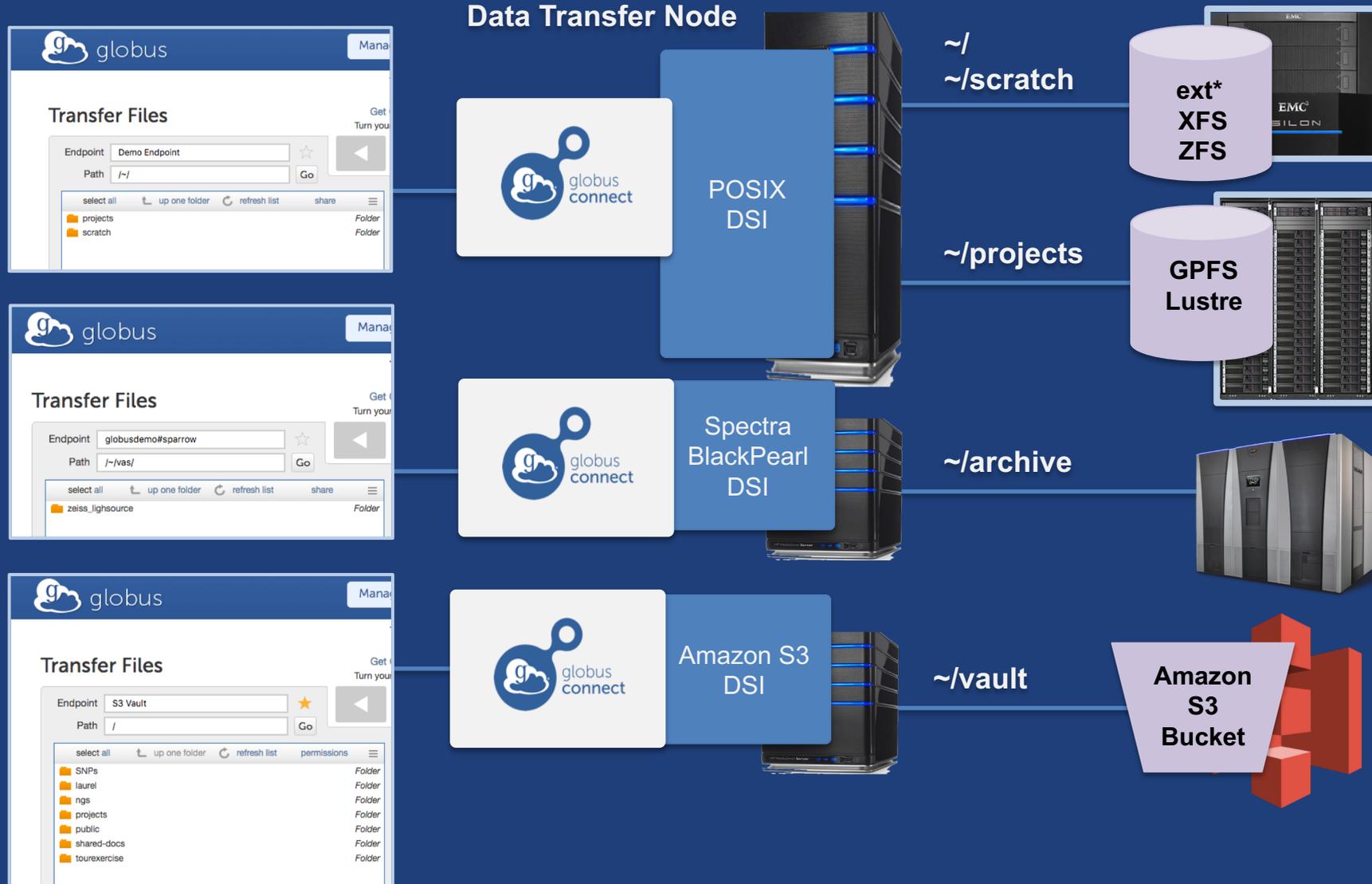


# Multi-endpoint configuration

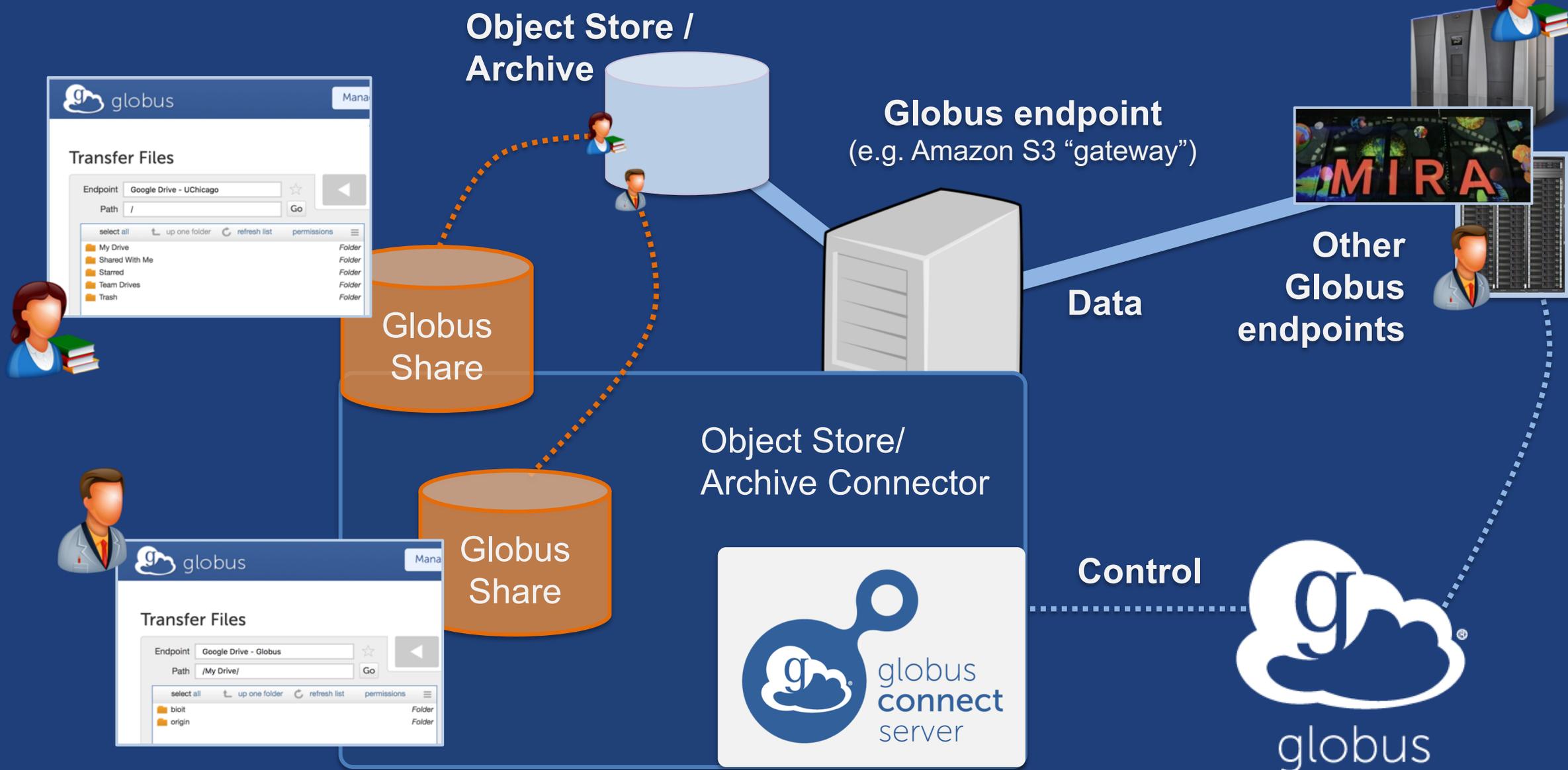




# Multi-endpoint configuration



# Deploying a premium connector gateway

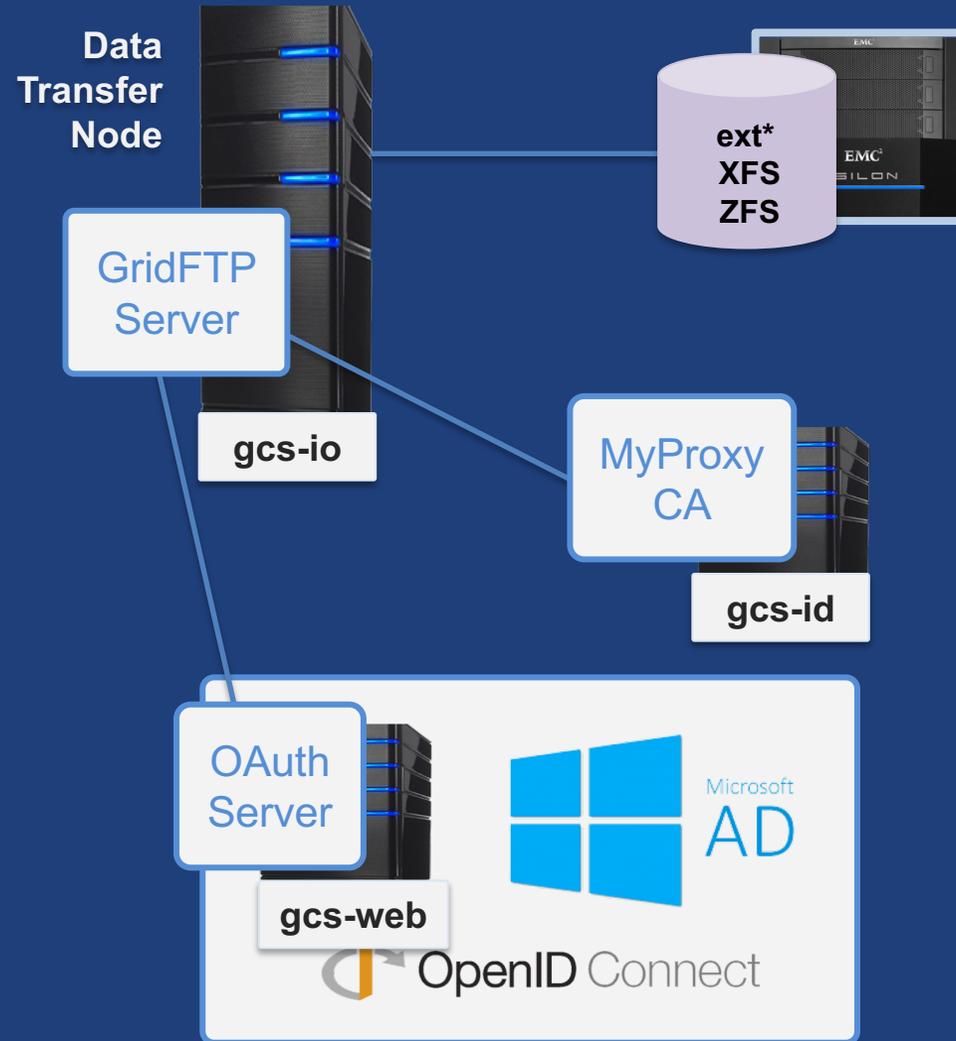




# Other Deployment Options



# Distributing Globus Connect Server components





# Example: Two-node DTN



On “primary” DTN node (34.20.29.57):  
/etc/globus-connect-server.conf  
[Endpoint] Name = **globus\_dtn**  
[MyProxy] Server = **34.20.29.57**



On other DTN nodes:  
/etc/globus-connect-server.conf  
[Endpoint] Name = **globus\_dtn**  
[MyProxy] Server = **34.20.29.57**



# Open Discussion