Introduction to Globus for New Users SaaS for Research Data Management

Vas Vasiliadis
vas@uchicago.edu

NC State – March 27, 2018

globus

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

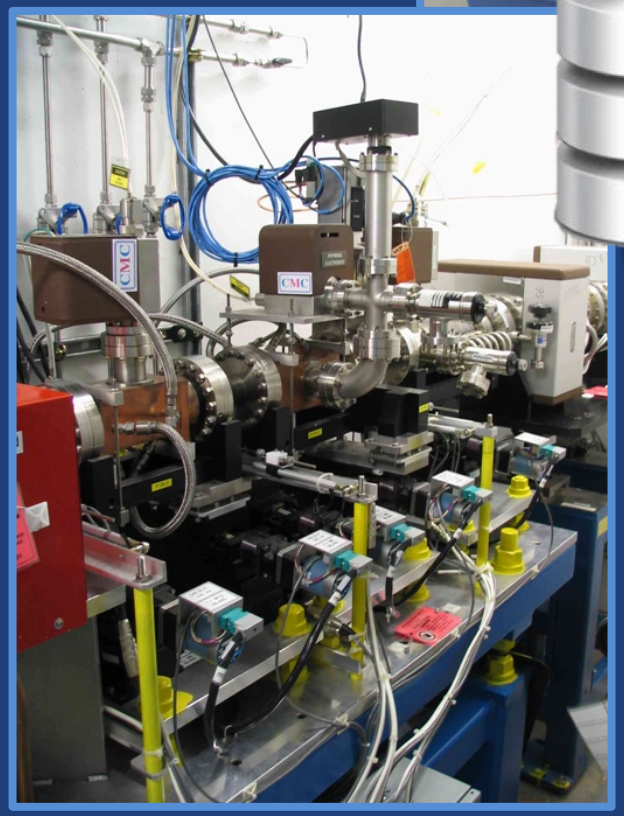
Move results to laptop, department, lab, …

Bridge to national cyberinfrastructure

Move datasets to supercomputer, national facility

Move results to campus (…)

Bridge to instruments



Pre-processed Data

Analysis store

globus connect

globus connect

Raw Source Data

High durability, low cost store

globus connect

Amazon Glacier

Bridge to collaborators

XSEDE
Extreme Science and Engineering Discovery Environment

EC2

Jetstream

NERSC

globus connect

CORNELL UNIVERSITY FOUNDED A.D. 1865

External Campus Storage

NC STATE UNIVERSITY

Public/Private Cloud stores

globus connect

red cloud

ceph

openstack

Google Drive

amazon web services S3

Bridge to community/public

globus connect

Project Repositories,
Replication Stores

globus connect

XSEDE
Extreme Science and Engineering Discovery Environment

EC2

Jetstream

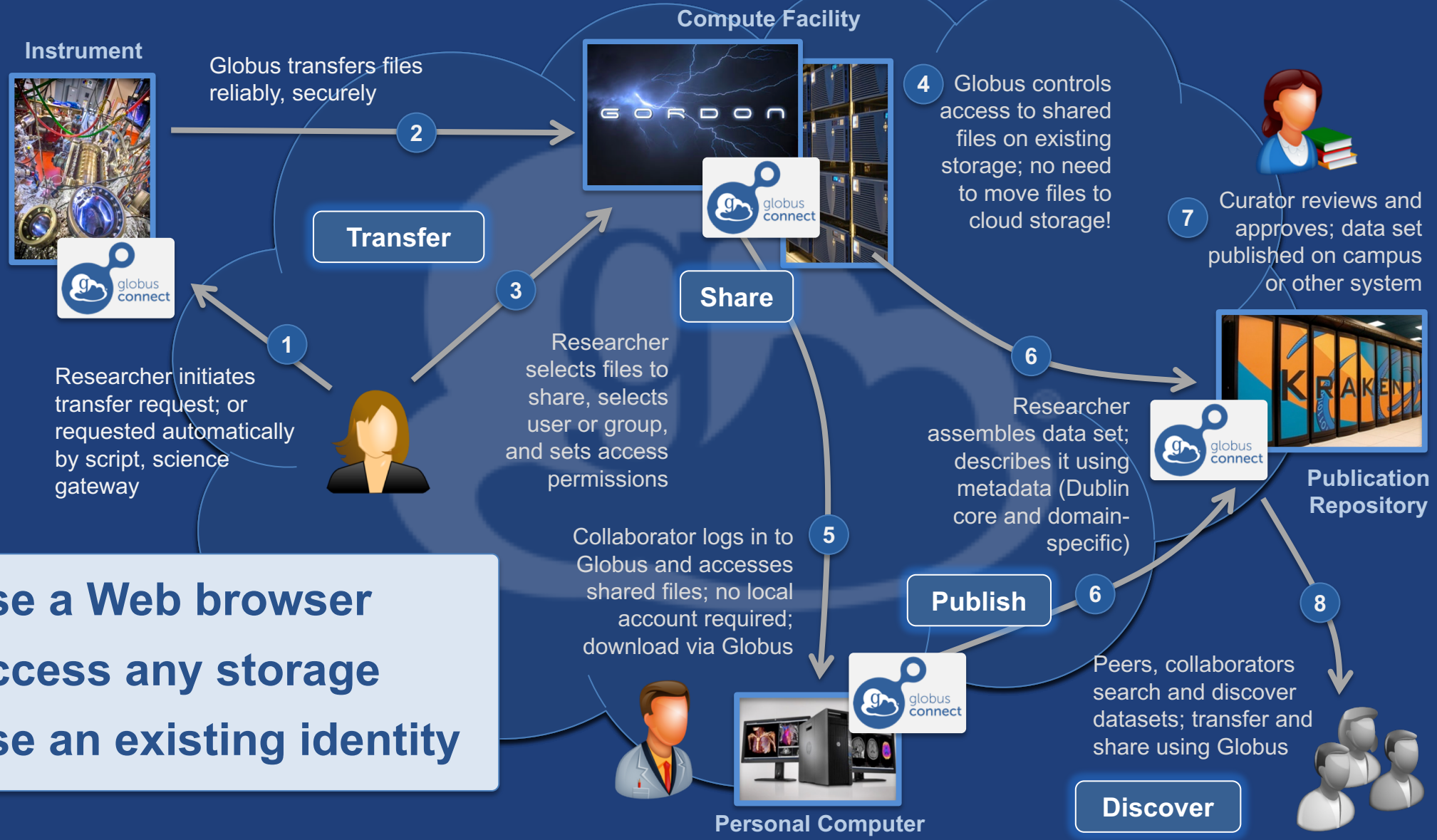
NERSC

Public Repositories

globus connect

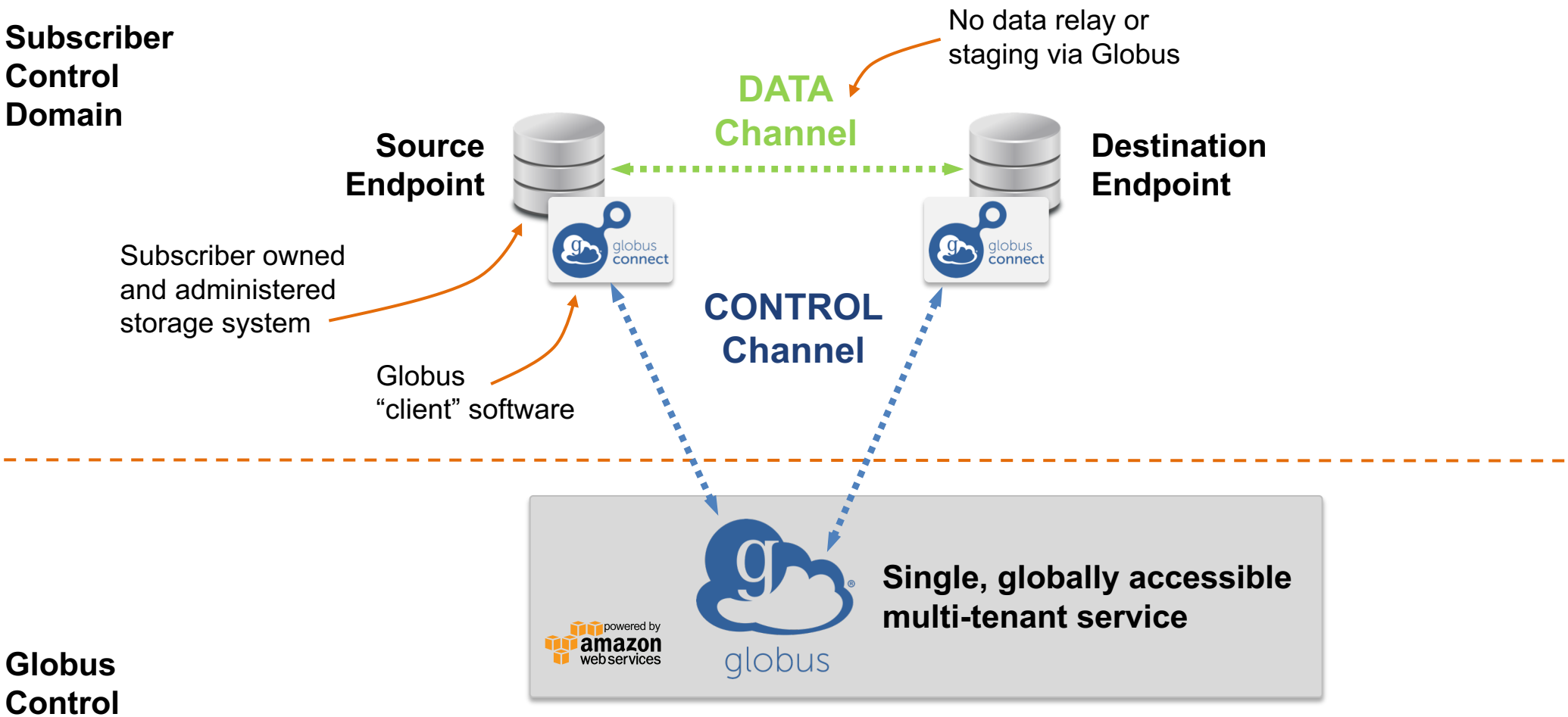
globus connect

Globus SaaS: Research data lifecycle

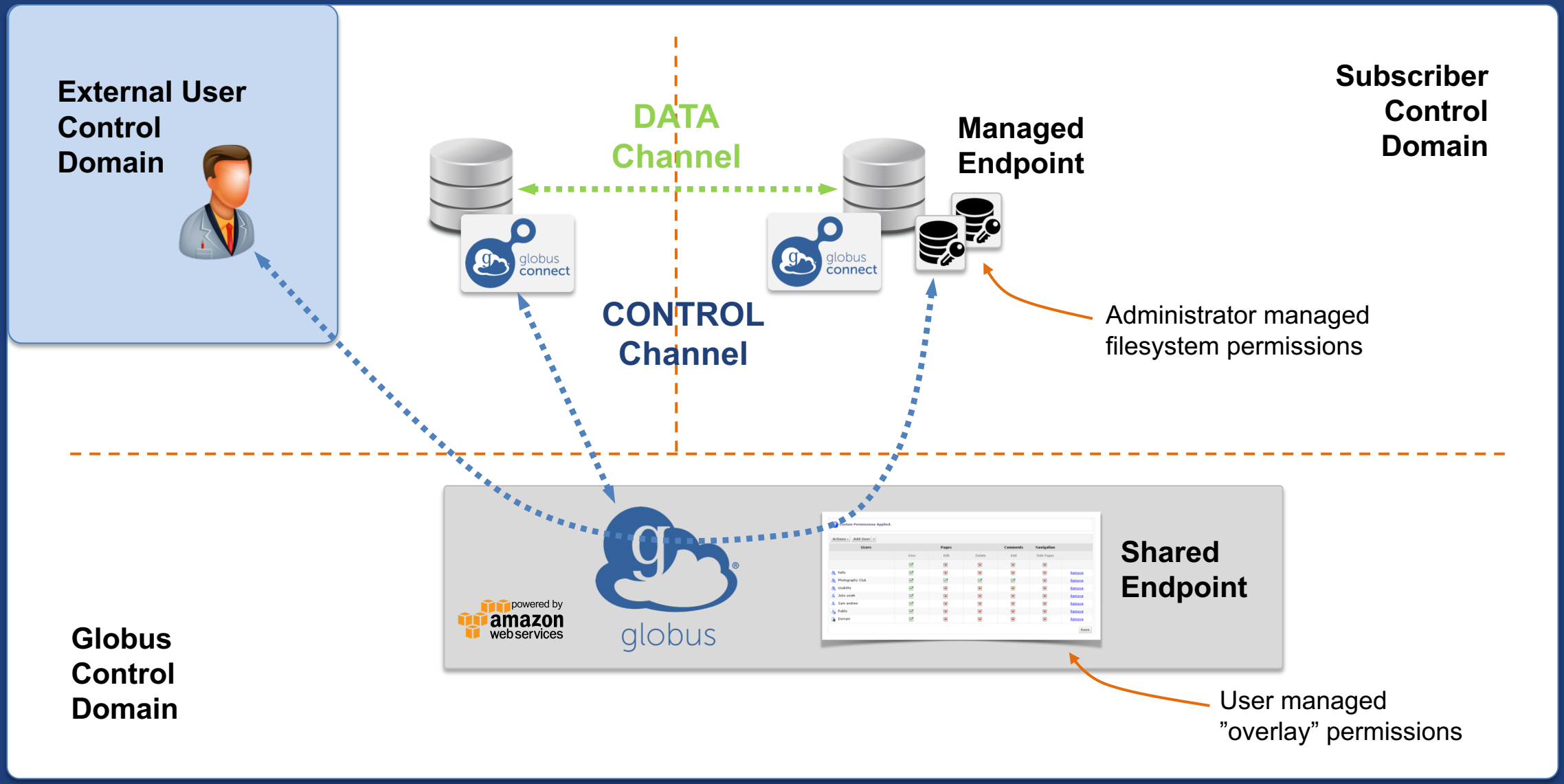


- **Use a Web browser**
- **Access any storage**
- **Use an existing identity**

Conceptual architecture: Hybrid SaaS



Conceptual architecture: Sharing



Why use Globus?

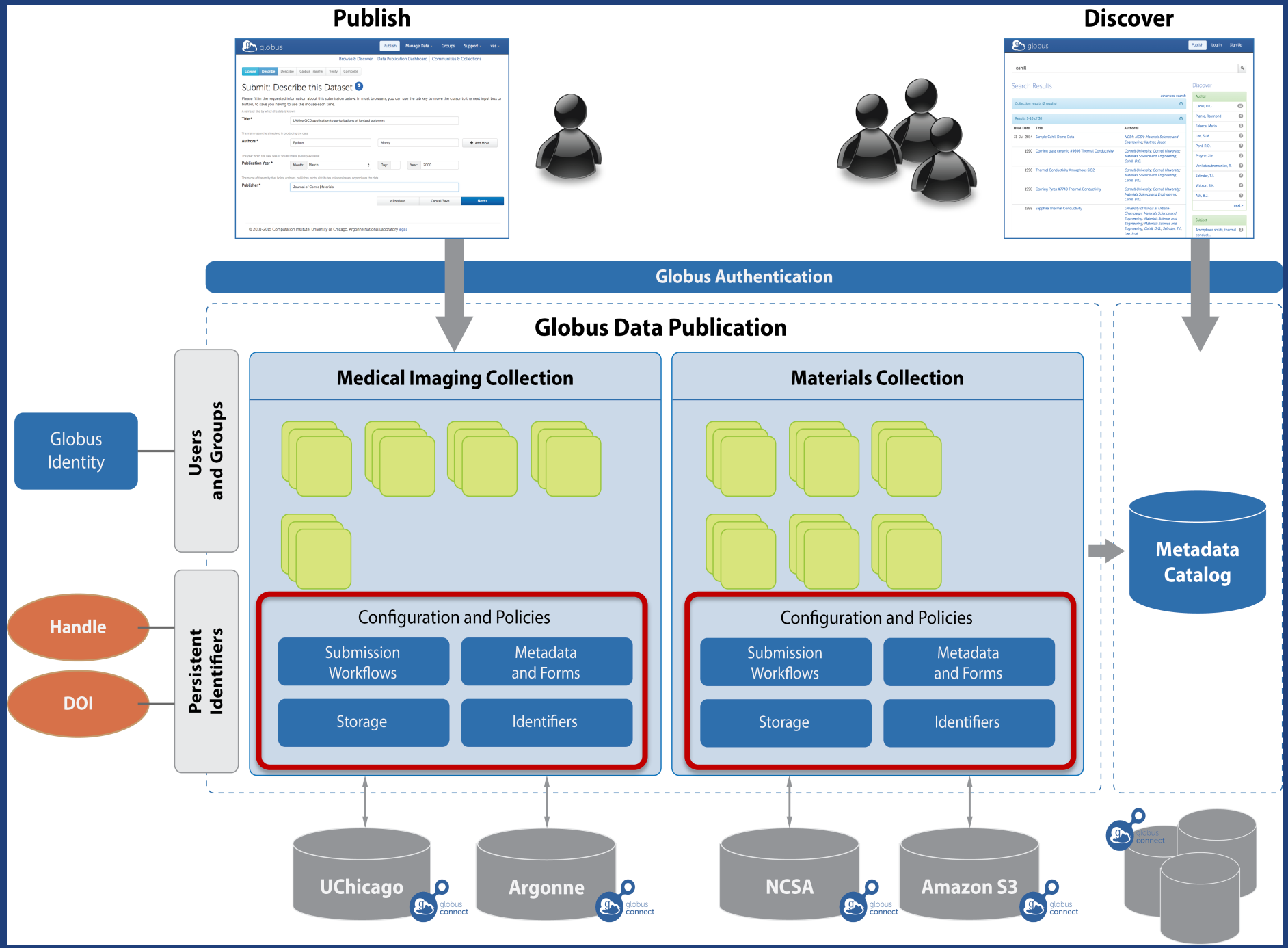
- **Simplicity**
  - Consistent UI across systems
  - Easy access to collaborators
- **Reliability and performance**
  - “Fire-and-forget” file transfer
  - Maximized WAN throughput
- **Operational efficiency**
  - Low overhead SaaS model
  - Highly automatable: CLI, RESTful API
- **Access to a large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

Demonstration
File Transfer
File Sharing
Group Management

Data Publication and Discovery

https://publish.globus.org

Globus data publication framework


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

Demonstration
Data Publication

How can I use Globus on my computer?

…makes your storage system a Globus endpoint

Globus Connect Personal

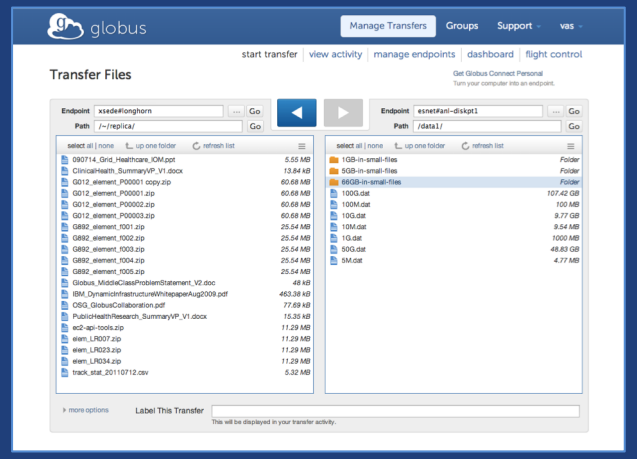
- **Installers do not require admin access**
- **Zero configuration; auto updating**
- **Handles NATs**

How can I integrate Globus into my research workflows?

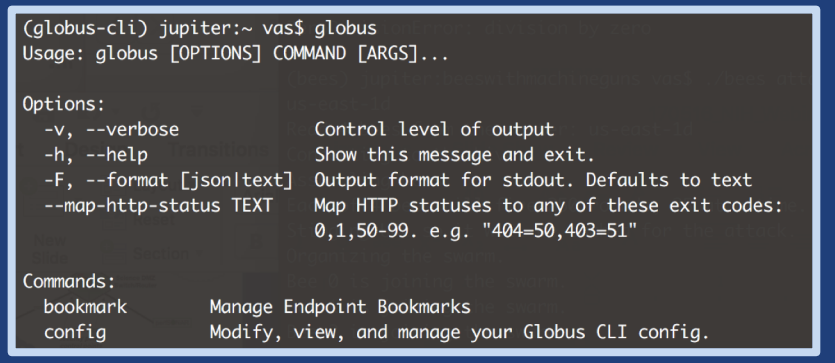
Globus serves as…

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

Use(r)-appropriate interfaces



Web



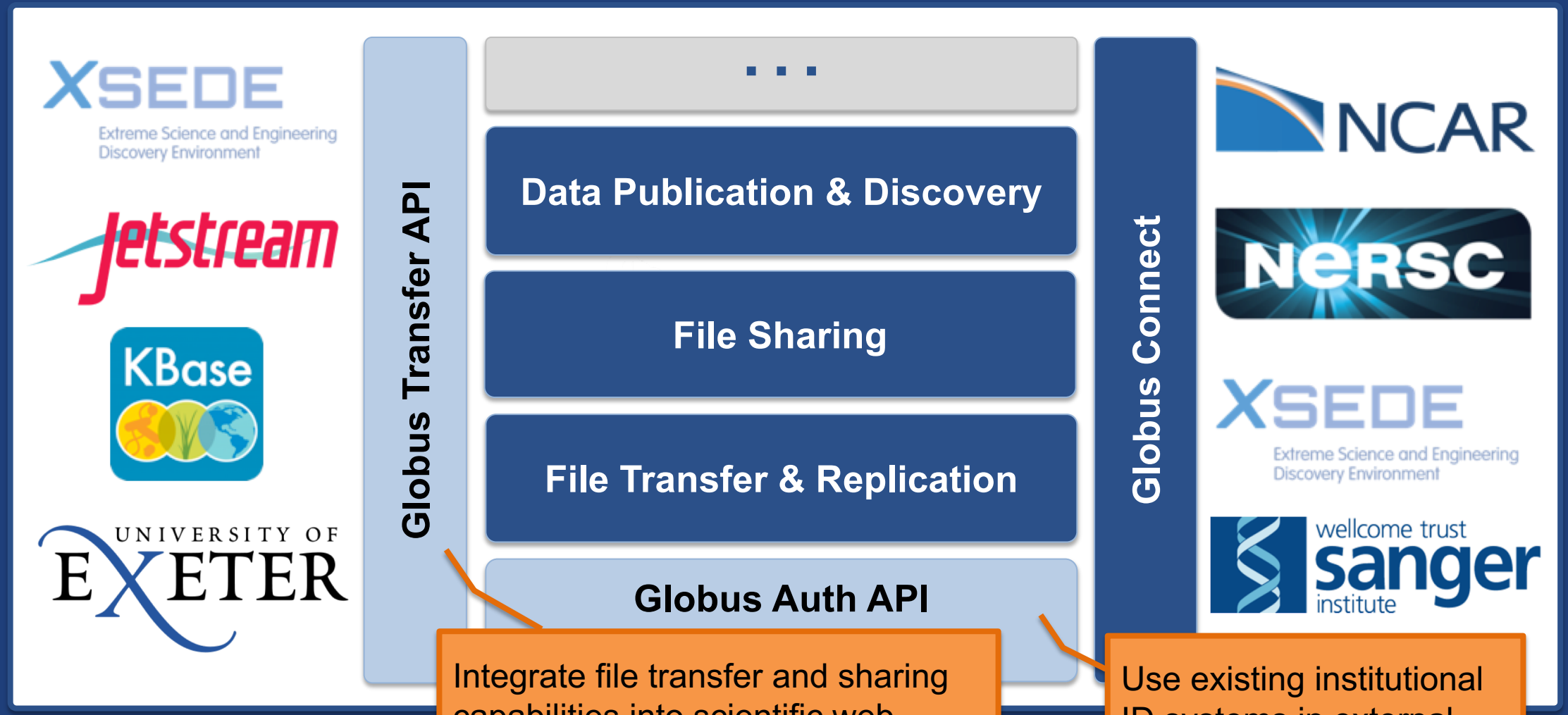
CLI

Globus service

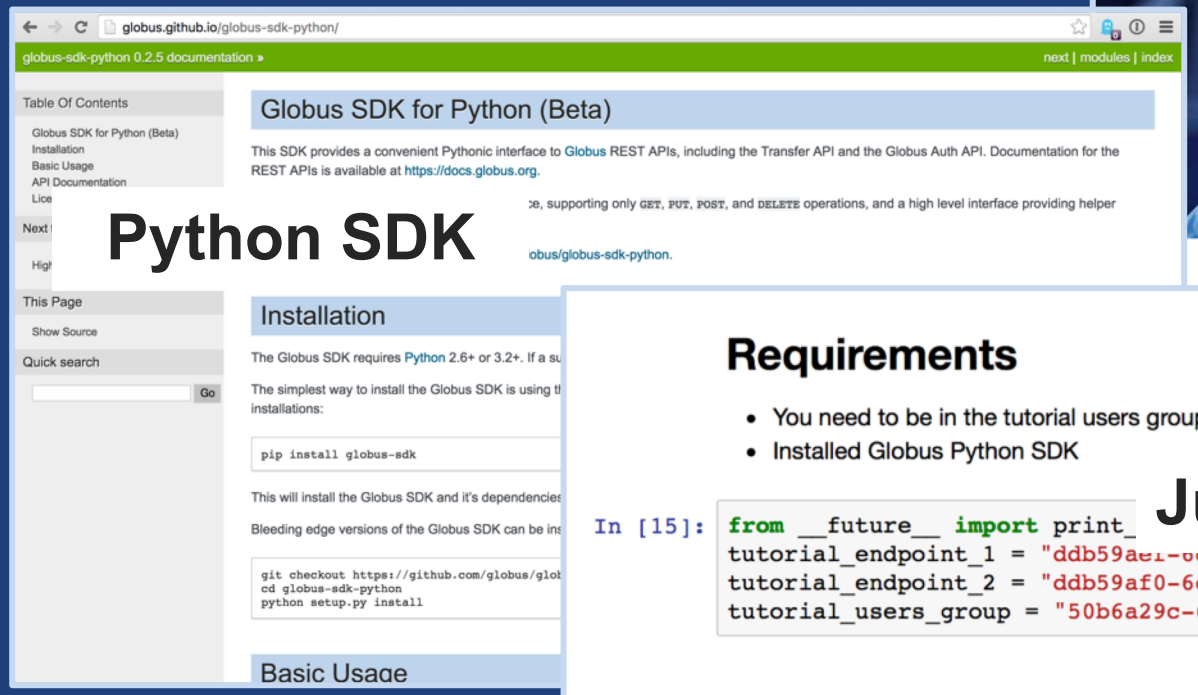
```
GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
…
```

Rest API

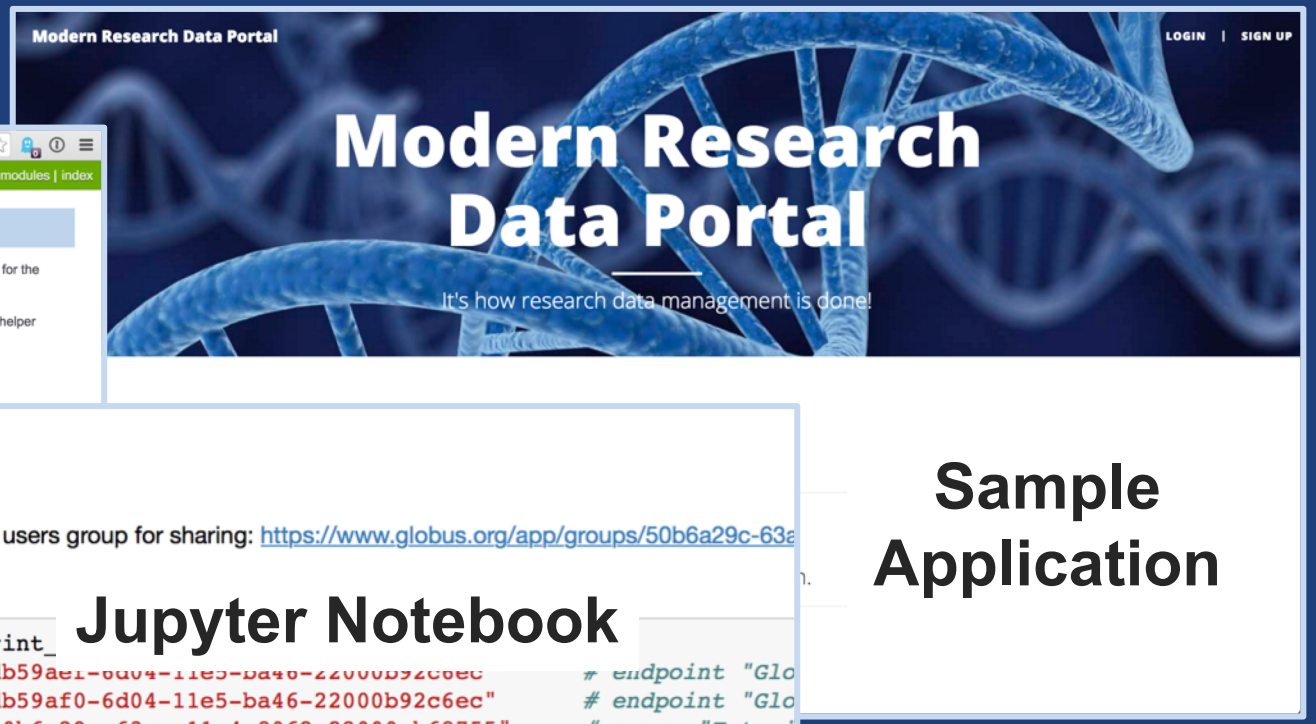
Globus as PaaS



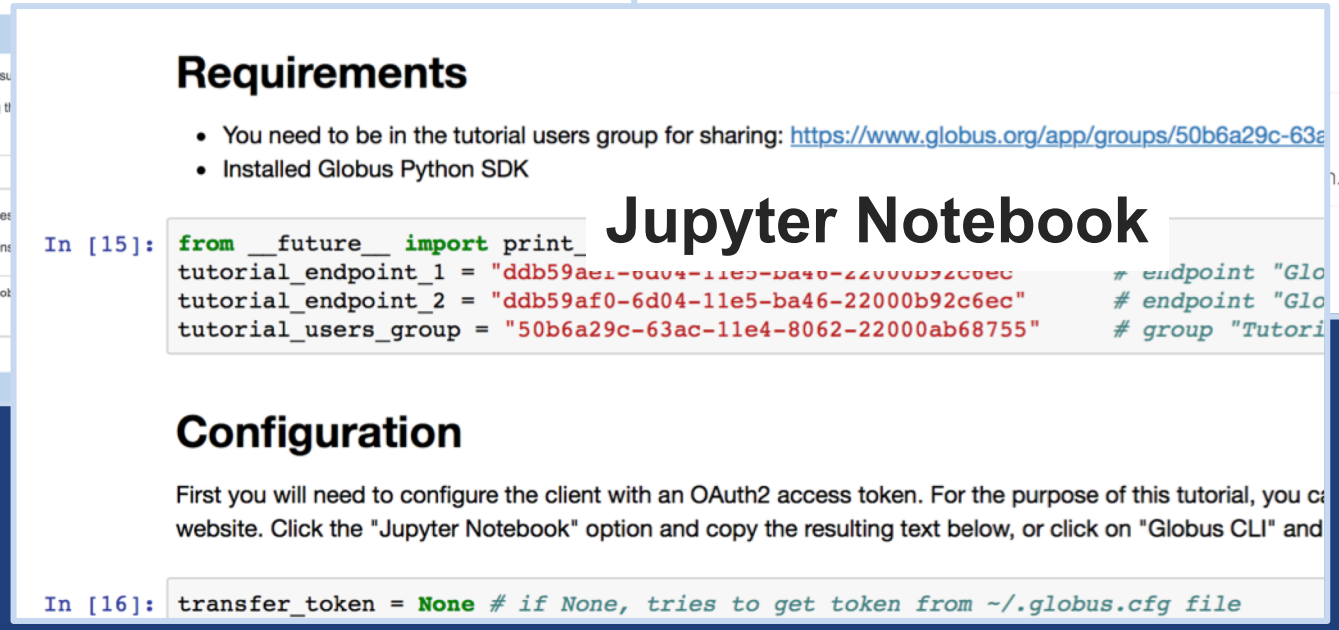
Globus PaaS developer resources



Python SDK



Sample Application



Jupyter Notebook

docs.globus.org/api

github.com/globus

Thank you to our sponsors…

U.S. DEPARTMENT OF
ENERGY

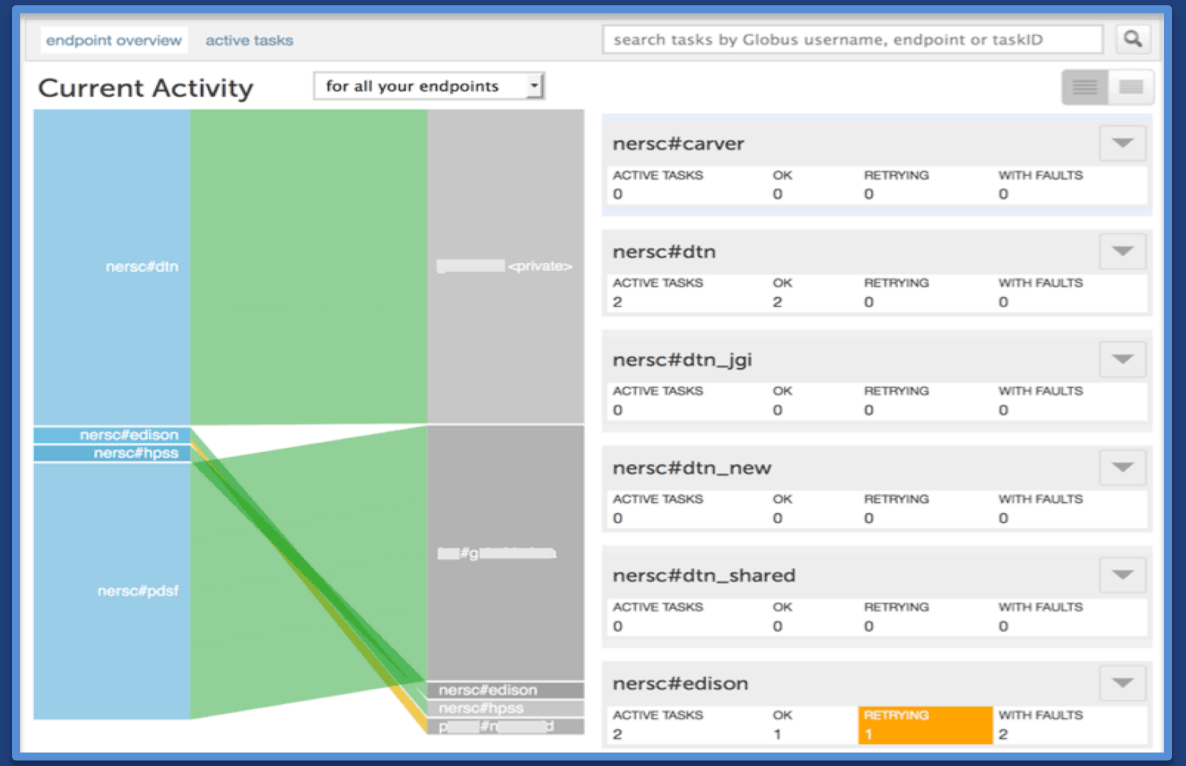
THE UNIVERSITY OF
CHICAGO

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

Argonne
NATIONAL LABORATORY

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

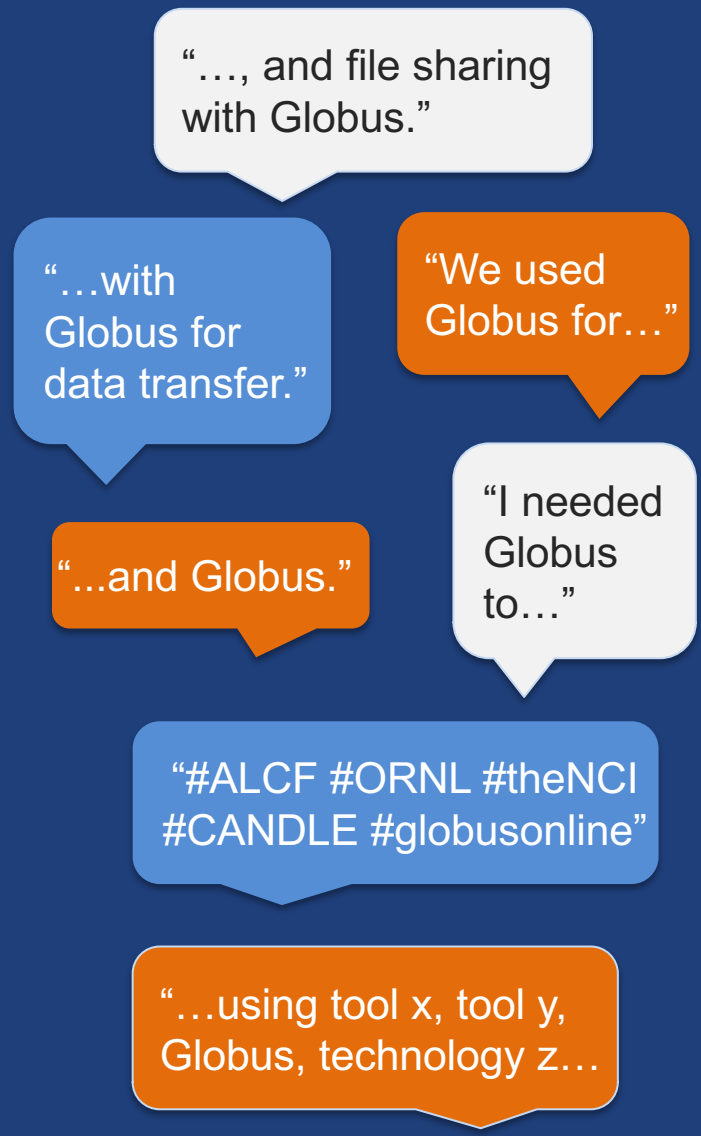
Our supporters

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**

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



Managing Globus Endpoints
Globus for System Administrators

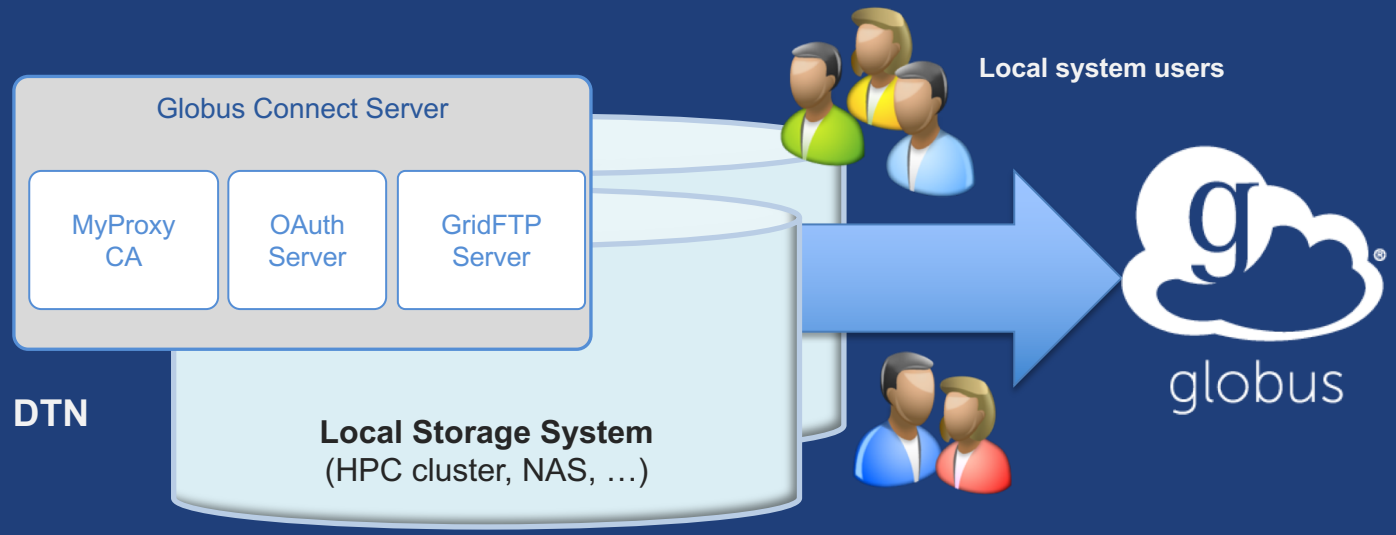
Vas Vasiliadis
vas@uchicago.edu

NC State – March 27, 2018

globus

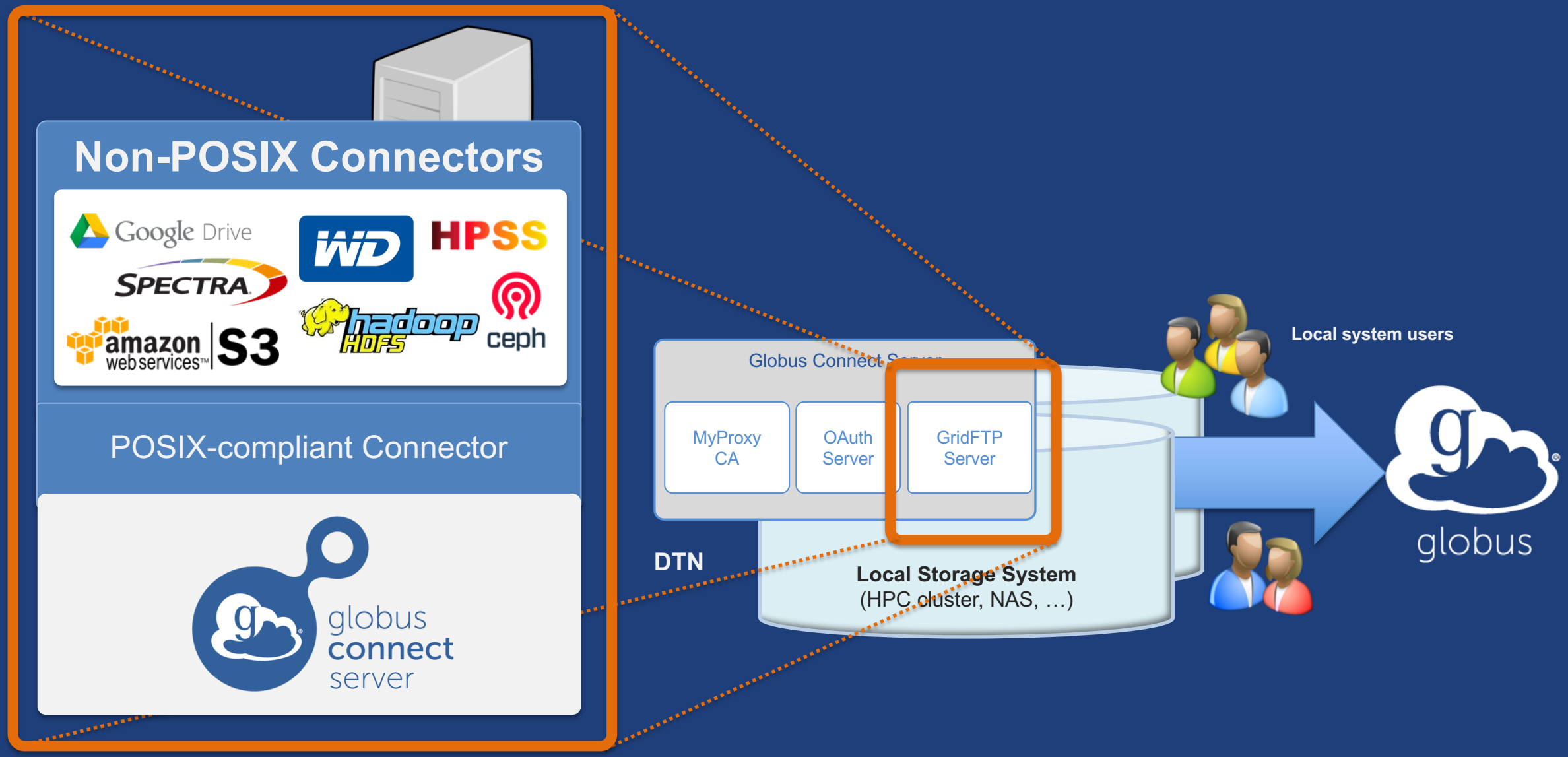
Globus Connect Server

- **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/

Globus Connect Server



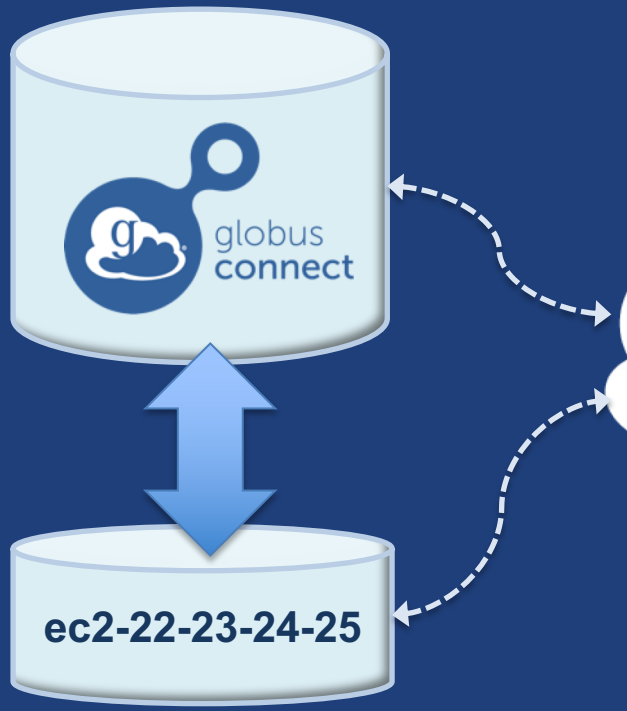
Creating a Globus endpoint on your server

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

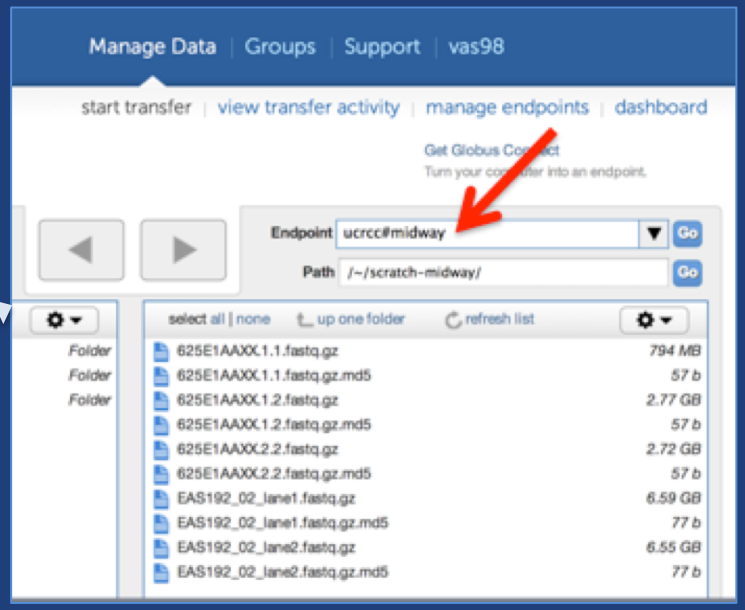
What we are going to do:

1 Install Globus Connect Server

- Access server as user "**campusadmin**"
- Update repo
- Install package
- Setup Globus Connect Server



2 Log into Globus



3 Access the newly created endpoint (as user 'researcher')

4 Transfer a file

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 -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 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

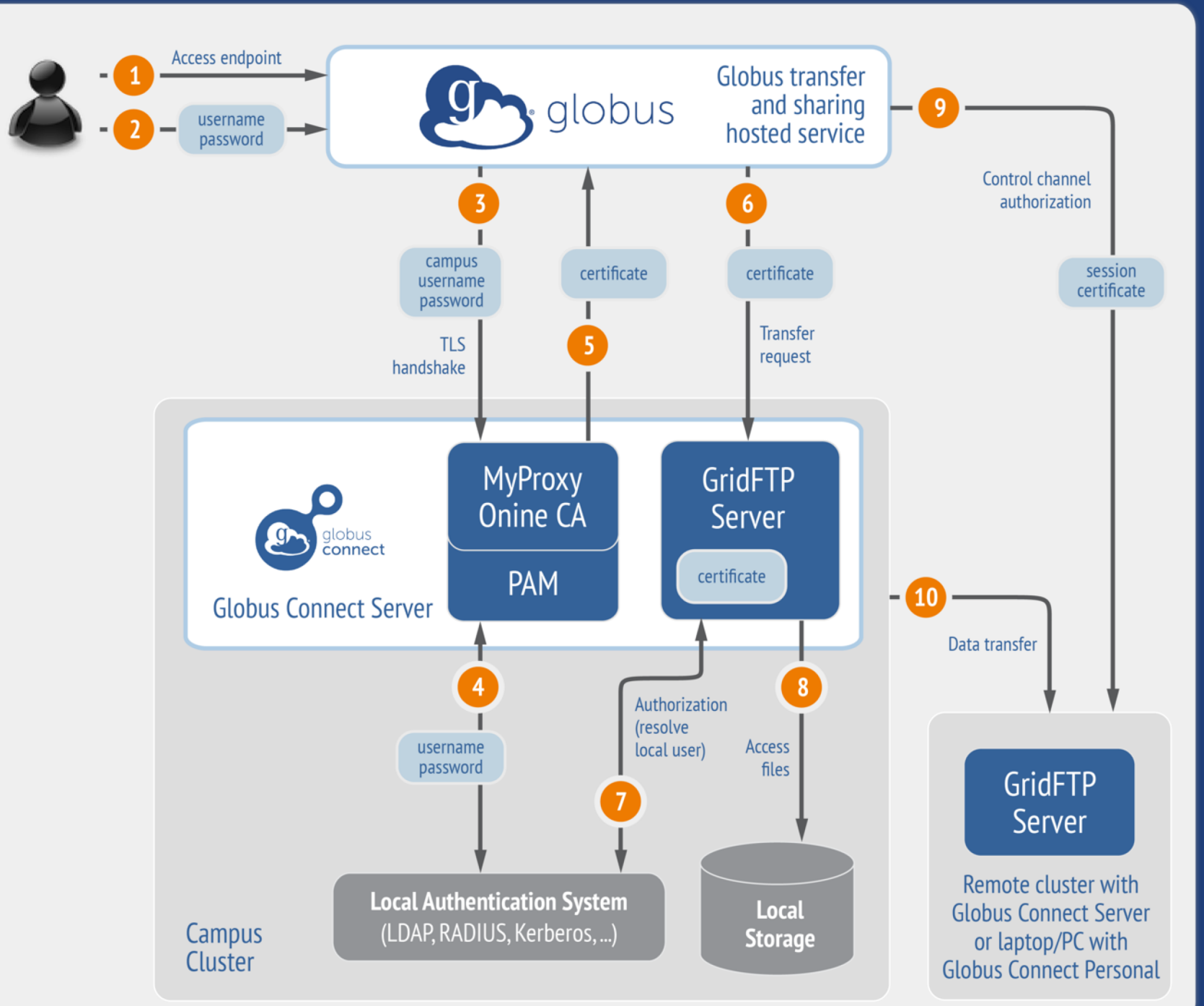
- **Restrict paths where users can create shared endpoints**
- **Use `SharingRestrictPaths` to customize**
  - Same syntax as `RestrictPaths`
- **e.g. Full access to home directory, deny hidden files:**
  - `SharingRestrictPaths = RW~,N~/.*`
- **e.g. Full access to public folder under home directory:**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

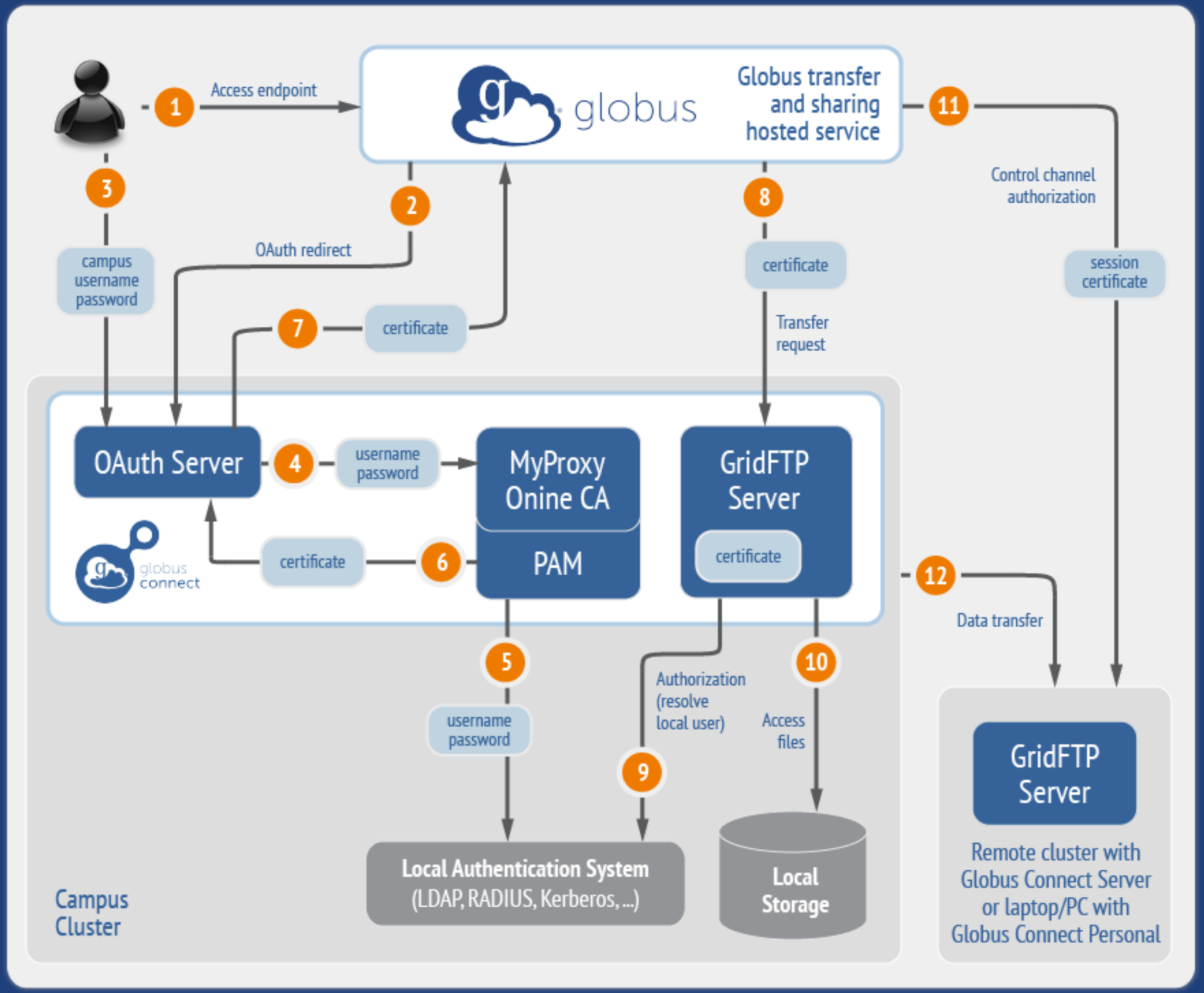
Endpoint activation using MyProxy



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

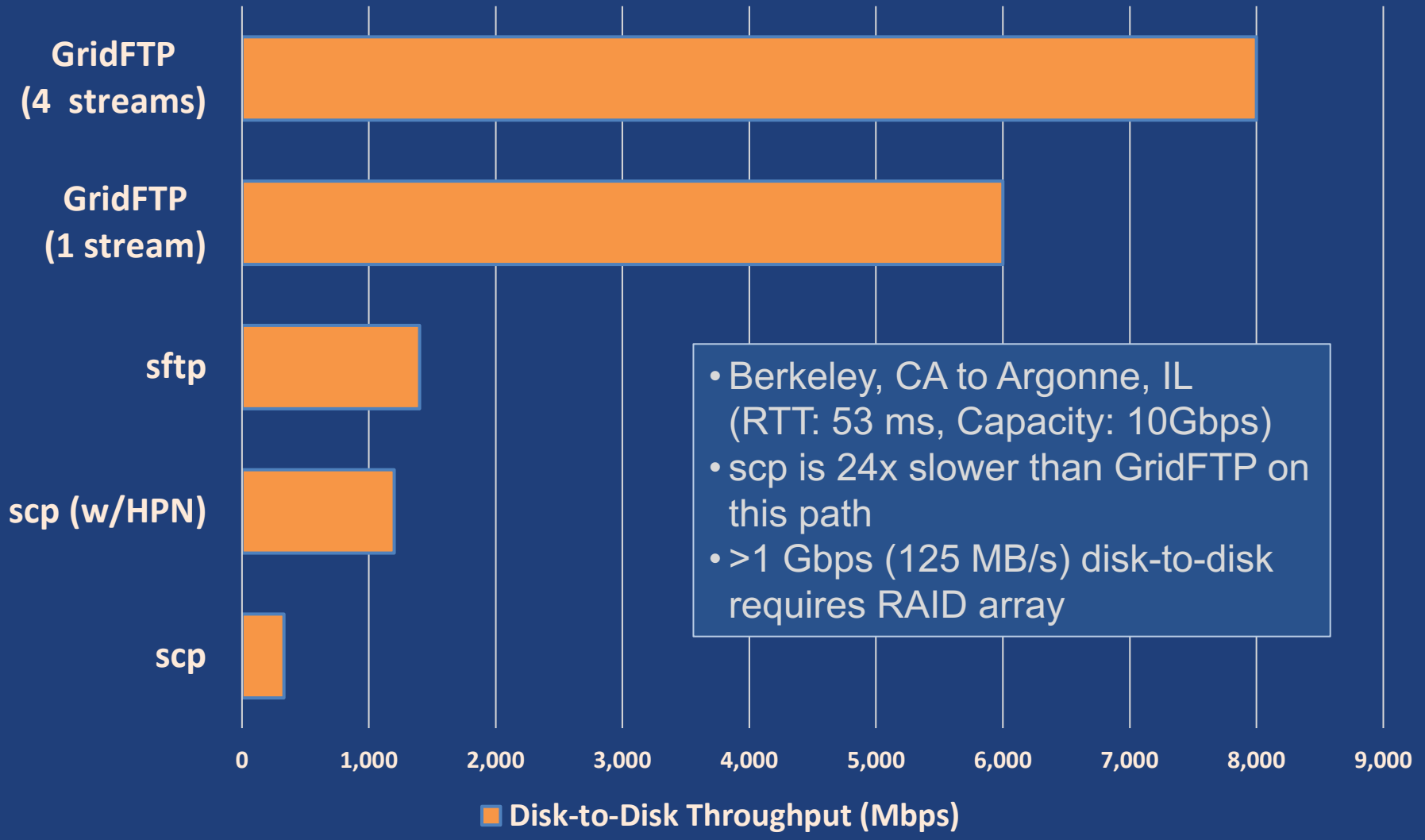
Balance: performance - reliability

- **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**

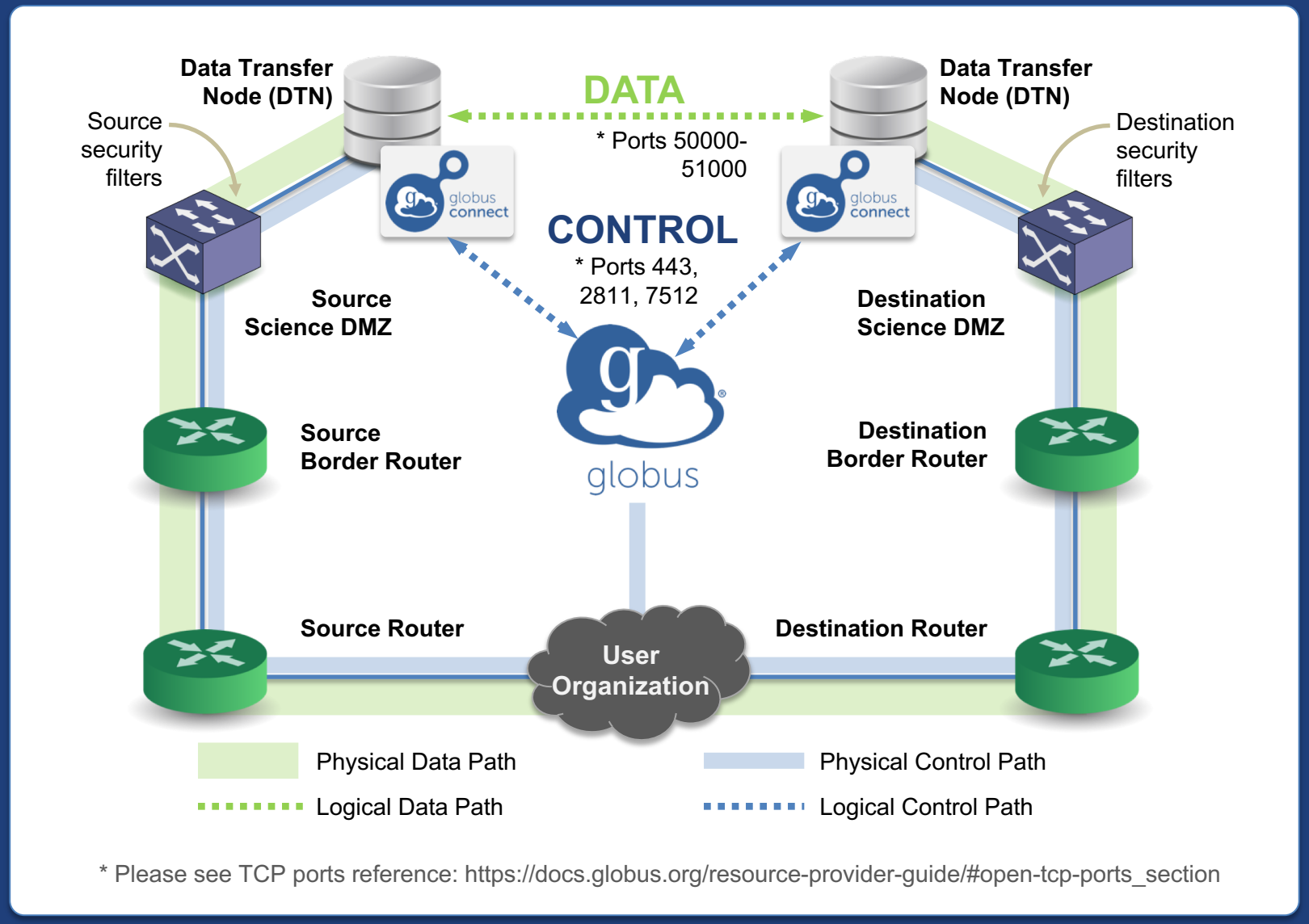
Disk-to-Disk Throughput: ESnet Testing



Source: ESnet (2016)

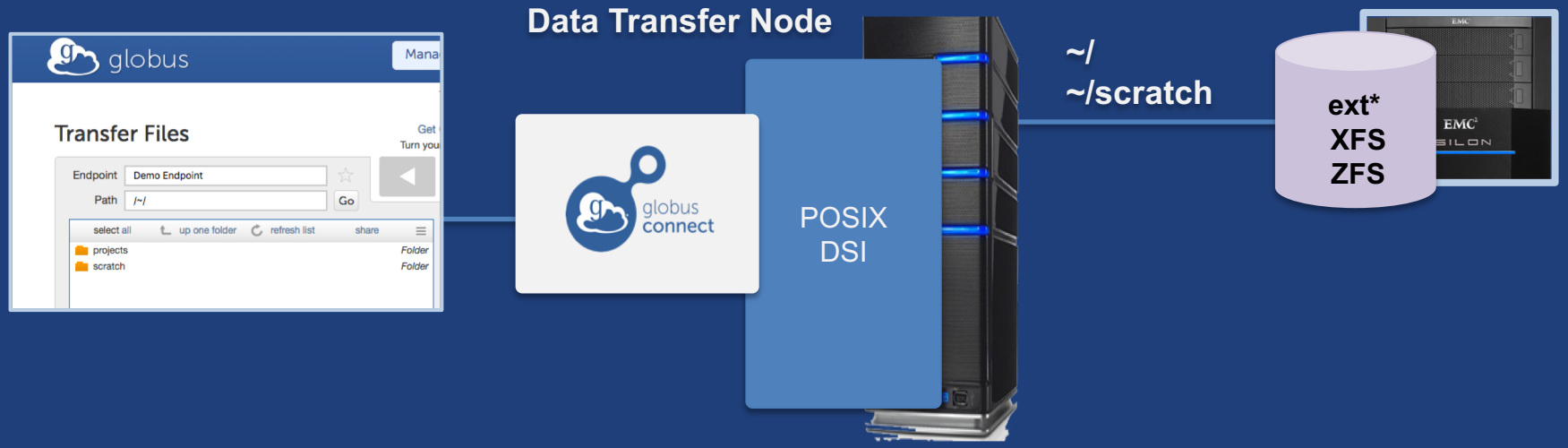
Deployment Scenarios

Best practice network configuration

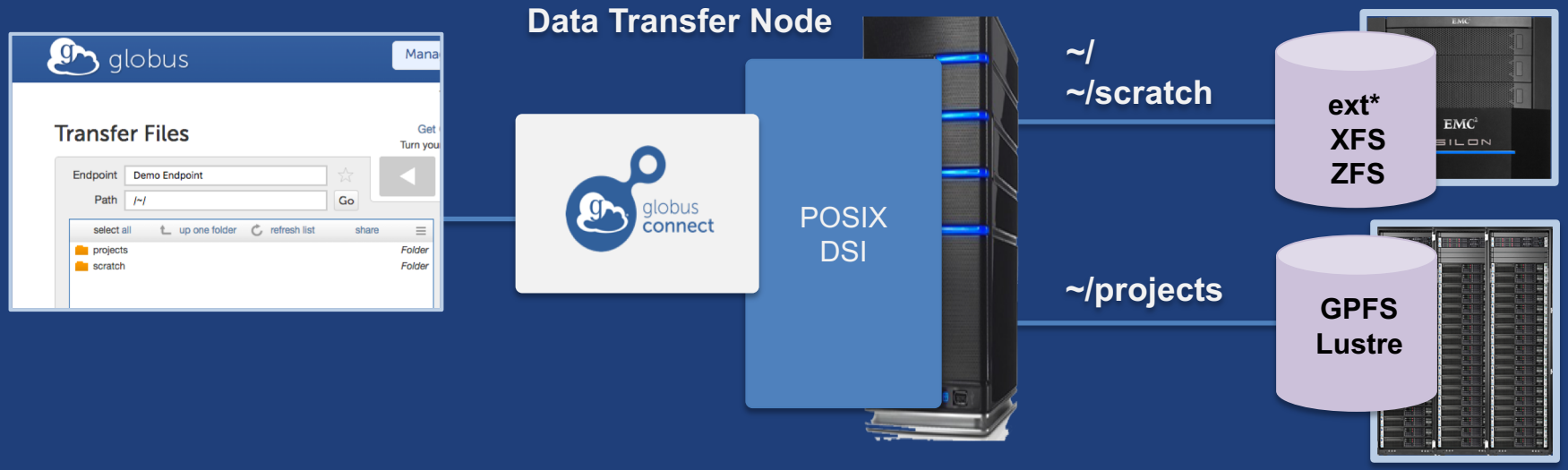


* Please see TCP ports reference: https://docs.globus.org/resource-provider-guide/#open-tcp-ports_section

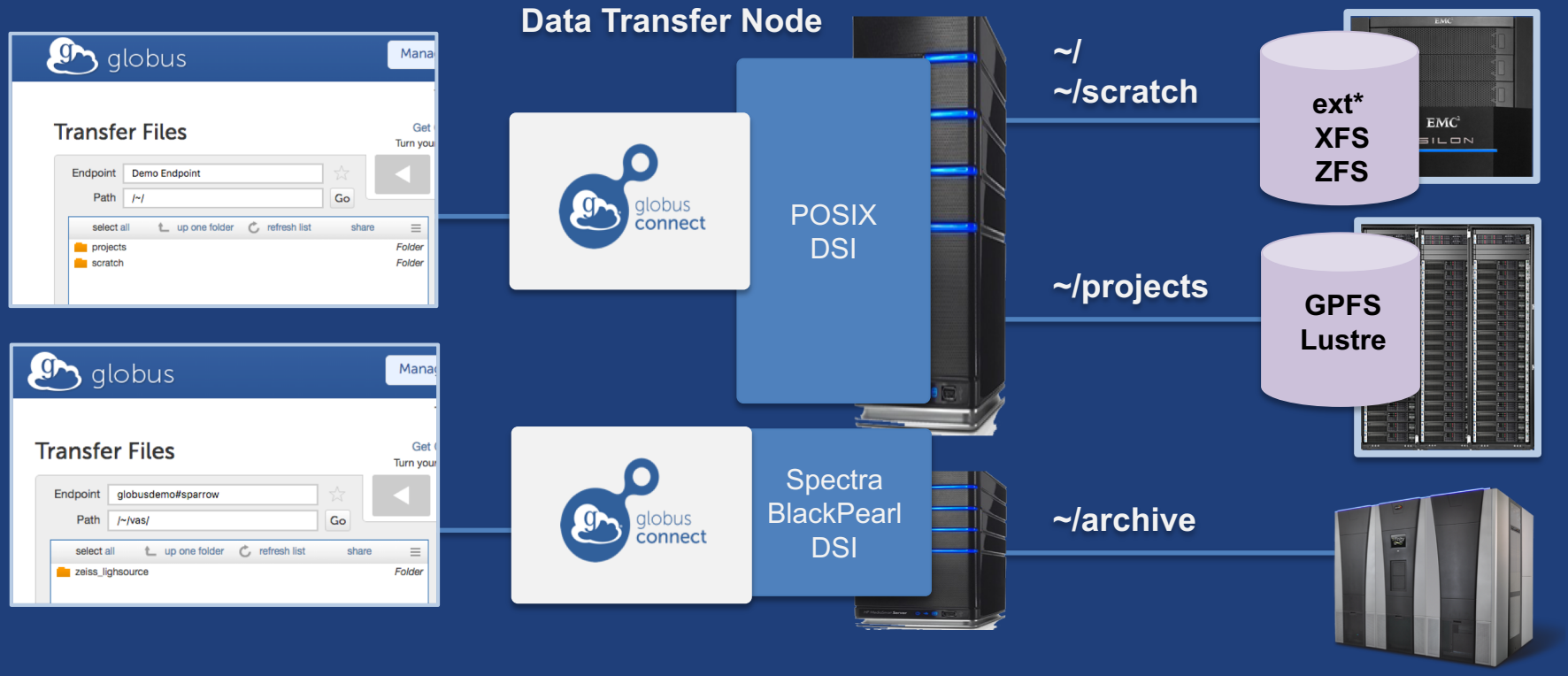
Multi-endpoint configuration



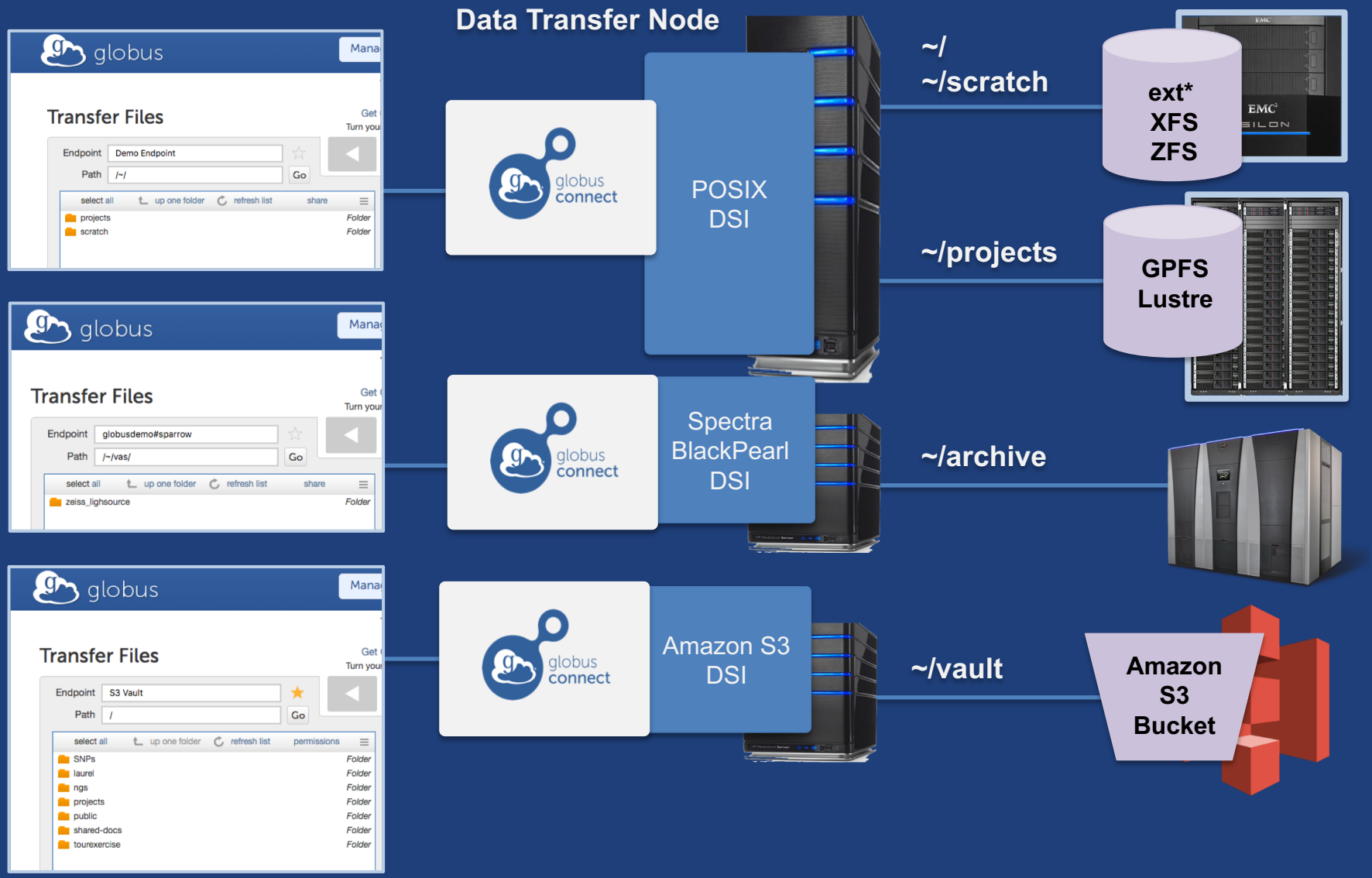
Multi-endpoint configuration



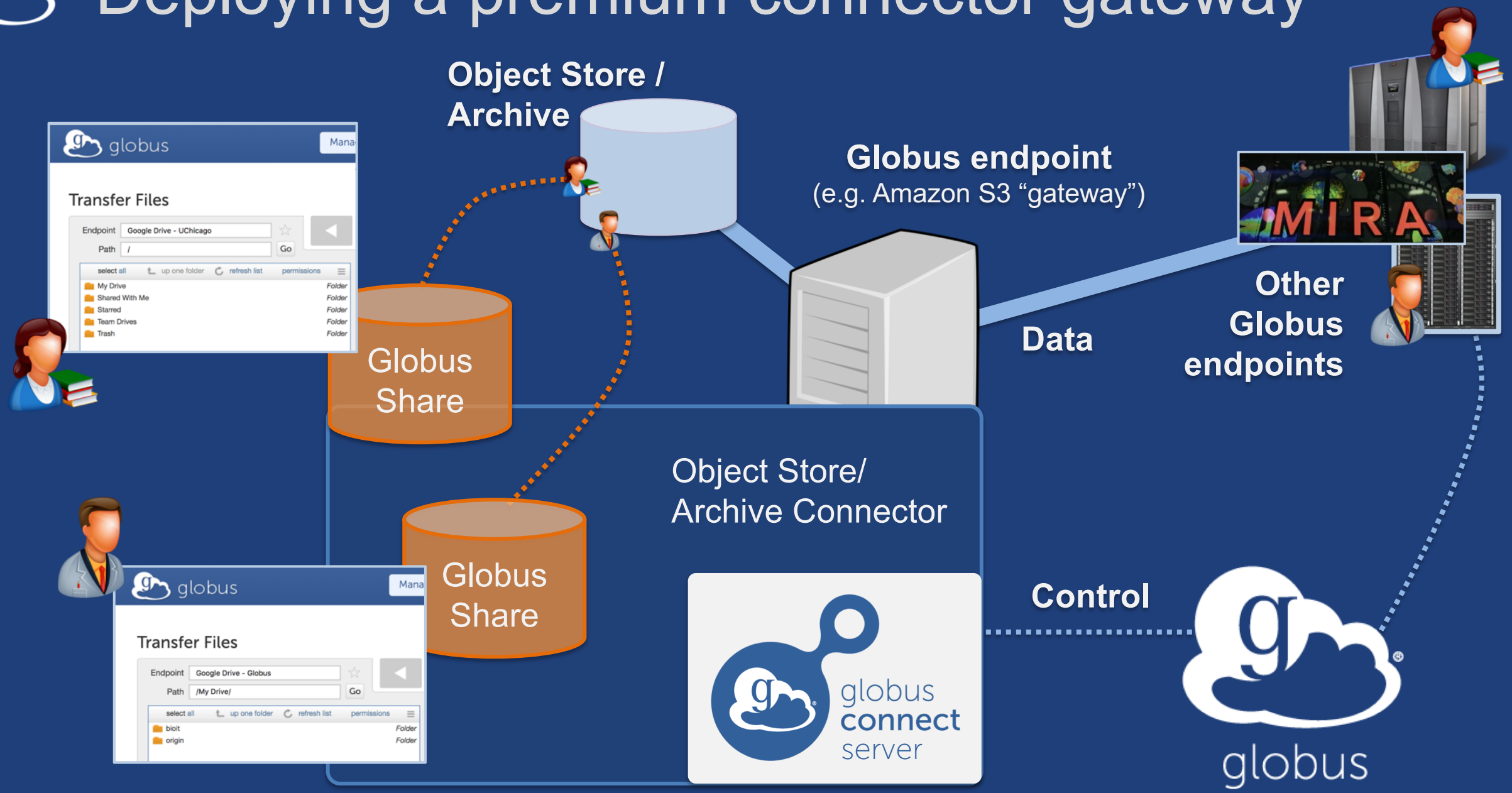
Multi-endpoint configuration



Multi-endpoint configuration

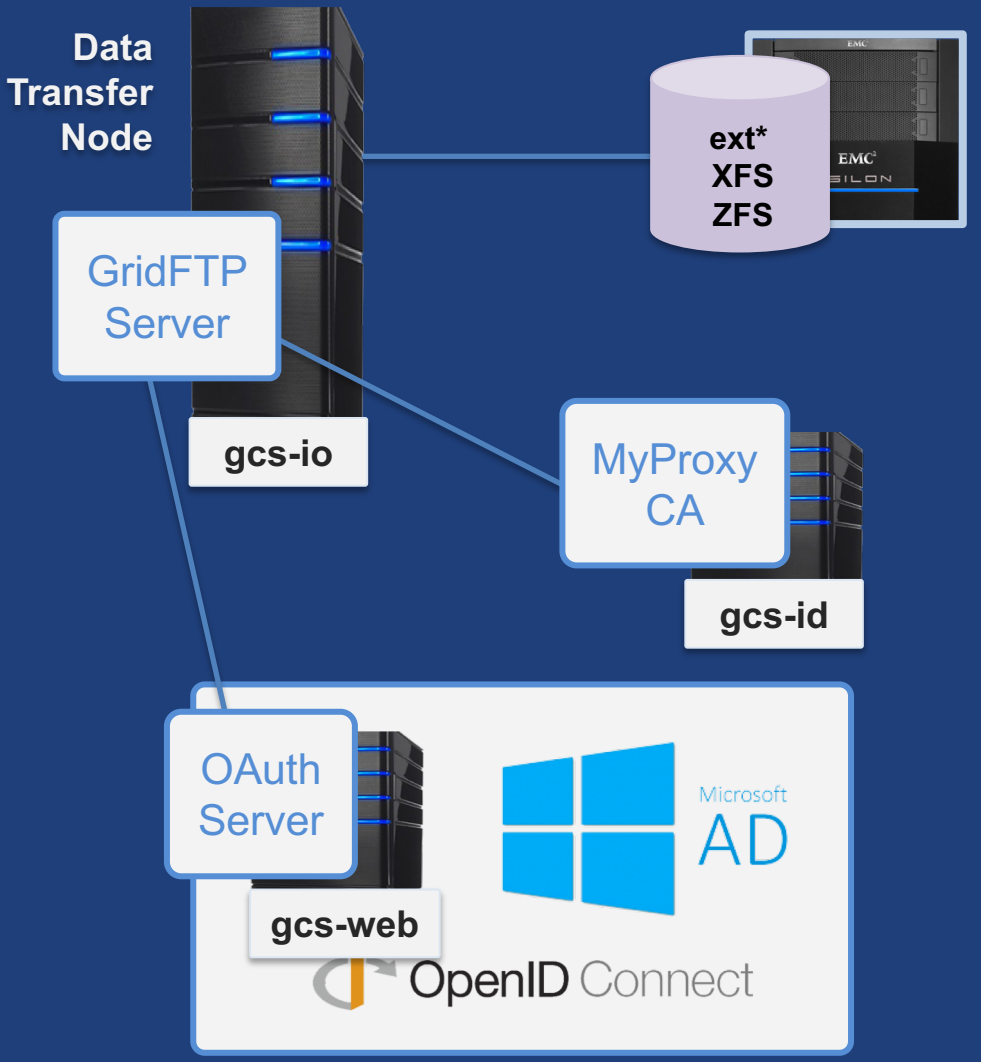


Deploying a premium connector gateway



Other Deployment Options

Distributing Globus Connect Server components



Example: Two-node DTN

-id
-io


On "primary" DTN node (34.20.29.57):

/etc/globus-connect-server.conf

```
[Endpoint] Name = globus_dtn
[MyProxy] Server = 34.20.29.57
```

-io


On other DTN nodes:

/etc/globus-connect-server.conf

```
[Endpoint] Name = globus_dtn
[MyProxy] Server = 34.20.29.57
```

Open Discussion