

Globus Research Data Management: Introduction and Service Overview

Steve Tuecke
tuecke@uchicago.edu

Vas Vasiliadis
vas@uchicago.edu



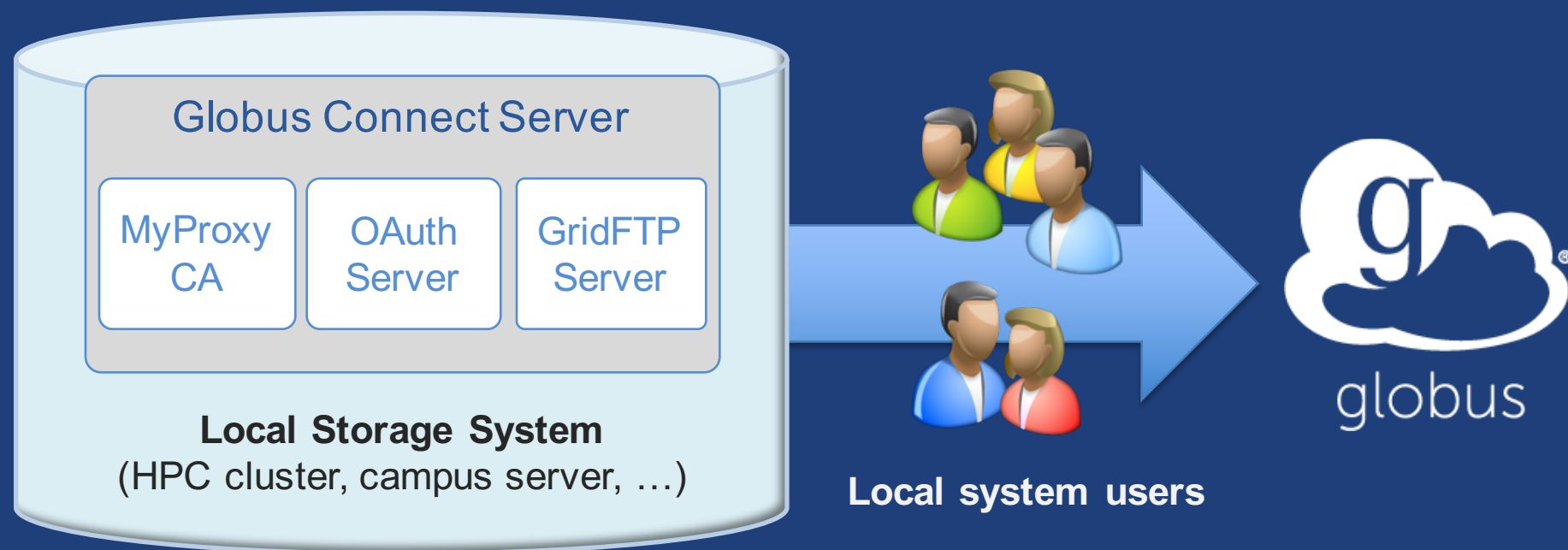


Agenda

- **Research data management challenges**
- **Globus: a high-level flyover**
- **File Transfer and Sharing: Accelerating and streamlining collaboration**
- **Data Publication: Enhancing reproducibility and discoverability**
- **Our sustainability challenge**
- **Globus campus deployment & intergation**
- **Deployment best practices: the Science DMZ**
- **Leveraging the Globus platform**



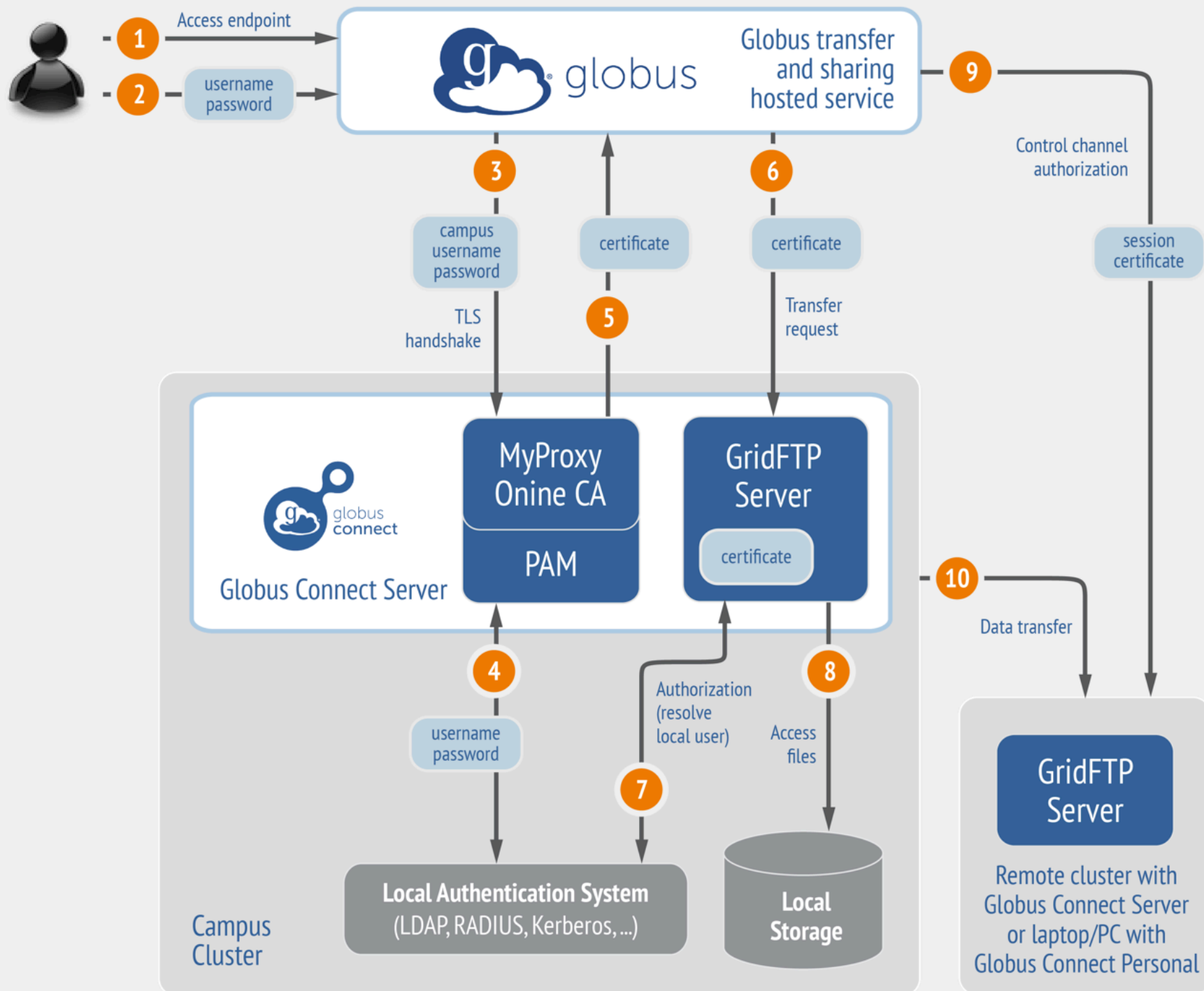
Globus Connect Server



- **Create endpoint in minutes; no complex software install**
- **Enable all users with local accounts to transfer files**
- **Native packages: RPMs and DEBs**

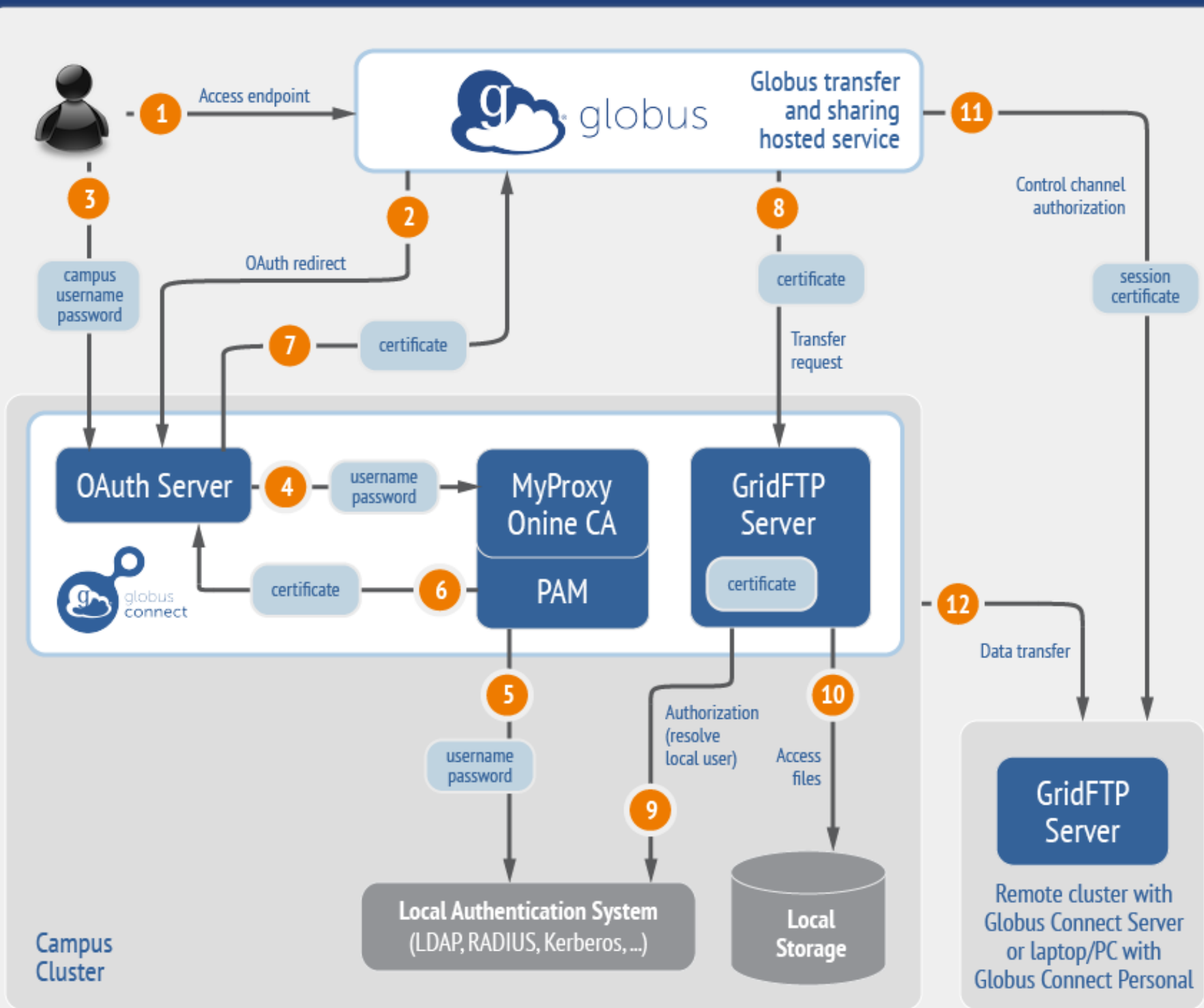


Endpoint activation using MyProxy





Endpoint activation using MyProxy OAuth





Standard package installation

1

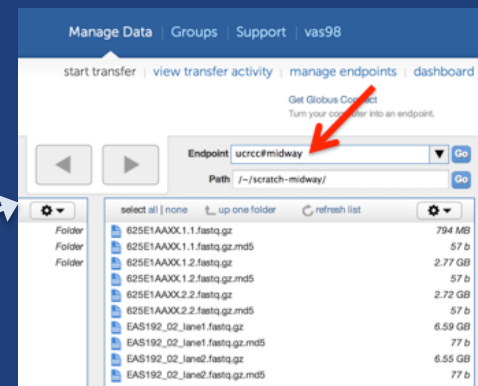
Install Globus Connect Server

- Access server as “campusadmin”
- Update package repos
- Install packages
- Setup Globus Connect Server



2

Log into Globus as “researcher”



3

Access newly created endpoint

4

Transfer a file



Globus Connect Server Demonstration



Exercise 4: Set up a Globus Connect Server endpoint and transfer files

- **Goal for this session: turn a storage resource into a Globus endpoint**
- **Each of you is provided with an Amazon EC2 server for this tutorial**



Step 1: Log into your host

- **Your slip of paper has the host information**
- **Log in as user 'campusadmin':**
`ssh campusadmin@<your-AWS-IP-address>`
(password: sc15globus)
- **NB: Please sudo su before continuing**
 - User 'campusadmin' has passwordless sudo privileges



Step 2: Install Globus Connect Server

‘Cheat sheet’: bit.ly/globus-sc15

```
$ 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 username/password here

You have a working Globus endpoint!



Step 3: Access your Globus endpoint

- **Go to Manage Data → Transfer Files**
- **Access the endpoint you just created**
 - Enter: `<username>#ec2-...` in Endpoint field
 - Log in as user “researcher” (pwd: sc15globus);
You should see the user’s home directory
- **Transfer files**
 - Between `esnet#???-diskpt1` and your endpoint



Configuring Globus Connect Server

- **Globus Connect Server configuration is stored in:**
 - `/etc/globus-connect-server.conf`
- **To enable configuration changes you must run:**
 - `globus-connect-server-setup`
- **“Rinse and repeat”**
- **NB: Please `sudo su` before continuing**



Configuration file walkthrough

- **Structure based on .ini format:**
[Section]
Option
- **Most common options to configure**
Name
Public
RestrictedPaths
Sharing
SharingRestrictedPaths
IdentityMethod (CILogon, OAuth)



Changing your endpoint name

- Edit `/etc/globus-connect-server.conf`
- Set `[Endpoint] Name = "dtn"`
- Run `globus-connect-server-setup`
 - Enter your username/password when prompted
- **Access the endpoint in your browser using the new endpoint name**
 - You may need to refresh your browser to see the new name in the endpoint list



Making your endpoint public

- Try to access the endpoint created by the person sitting next to you
- You will get the following message:
- ‘Could not find endpoint with name ‘dtn’ owned by user ‘<neighbor’s username>’



Making your endpoint public

- **Edit: `/etc/globus-connect-server.conf`**
- **Uncomment `[Endpoint] Public` option**
- **Replace `False` with `True`**
- **Run `globus-connect-server-setup`**
- **Try accessing your neighbor's endpoint:
you will be prompted for credentials...**
- **...you can access the endpoint as the
“researcher” user**



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~/.*`



Sharing Path Restriction

- **Further restrict the paths on which your users are allowed to 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`



Control sharing access to specific accounts

- **SharingStateDir** can be used to control sharing access to individual accounts
- **For instance, with**
`SharingStateDir = "/var/globus/sharing/$USER"` **user "bob" would be enabled for sharing only if a path exists with the name "/var/globus/sharing/bob/" and is writable by bob.**



Using MyProxy OAuth server

- **MyProxy without OAuth (we just did this!)**
 - Site passwords flow through Globus to site MyProxy server
 - Globus does not store passwords
 - Still a security concern for some sites
- **Web-based endpoint activation**
 - Sites run a MyProxy OAuth server
 - MyProxy OAuth server in Globus Connect Server
 - Users enter username/password only on site's webpage to access an endpoint
 - Globus gets short-term X.509 credential via OAuth protocol



Single Sign-On with InCommon/CILogon

- **Requirements**
 - Your organization's Shibboleth server must release the ePPN attribute to CILogon
 - Your local resource account names must match your institutional identity (InCommon ID)
- **Set AuthorizationMethod = CILogon in the Globus Connect Server configuration**
- **Set CILogonIdentityProvider = <your_institution_as_listed_in_CILogon_identity_provider_list>**
- **Add CILogon CA to your trustroots**
 - /var/lib/globus-connect-server/grid-security/certificates/
 - Visit ca.cilogon.org/downloads for certificates



Using a host certificate for GridFTP

- **You can use your GridFTP server with non-Globus clients**
 - Requires a host certificate, e.g. from OSG
- **Comment out**
 - `FetchCredentialFromRelay = True`
- **Set**
 - `CertificateFile =`
`<path_to_host_certificate>`
 - `KeyFile = <path_to_private`
`key_associated_with_host_certificate>`
 - `TrustedCertificateDirectory =`
`<path_to_trust_roots>`



Enable sharing on your endpoint

- Edit: `/etc/globus-connect-server.conf`
- Uncomment `[GridFTP] Sharing = True`
- Run `globus-connect-server-setup`
- Go to the Web UI Start Transfer page*
- Select the endpoint*
- Create shared endpoints and grant access to other Globus users*

* Note: Creation of shared endpoints requires a **Globus Provider** plan for the managed endpoint
Contact support@globus.org for a one-month free trial



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: Run the above command after deleting/re-creating endpoint**



Demonstration: Globus Command Line Interface (CLI)



Exercise: Globus CLI

1. Optional: Generate SSH key
2. Go to:
globus.org/account/ManageIdentities
3. Add SSH key to your Globus identity
4.

```
ssh <username>@cli.globusonline.org
```
5. Check on status of earlier transfer(s)
6. Optional: Transfer a file using the `transfer` command



Deployment Scenarios

- **Globus Connect Server components**
 - globus-connect-server-io, -id, -web
- **Default: -io and -id (no -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**



Firewall configuration

- **Allow inbound connections to port:**
 - 2811 (GridFTP control channel)
 - 7512 (MyProxy CA) or 443 (OAuth)
- **Allow inbound connections to ports 50000-51000 (GridFTP data channel)**
 - If transfers to/from this machine will happen only from/to a known set of endpoints (not common), you can restrict connections to this port range only from those machines
- **If firewall restricts outbound connections, allow outbound connections if source port is:**
 - 80, 2223 (used during installation/configuration)
 - 50000-51000 (GridFTP data channel)



Deployment Best Practice: Science DMZ



Researchers don't realize full benefits of existing IT infrastructure

- **Impedance mismatch between research computing systems and the WAN**
- **Network “misconfiguration” (10 x 1Gb/s links \neq 1 x 10Gb/s link)**
- **Indiscriminate security policies**
- **TCP: small amount of packet loss = huge difference in performance**



Science DMZ Components

- **“Friction free” network path**
- **Dedicated, high-performance data transfer nodes (DTNs)**
- **Performance measurement/test node**
- **User engagement and education**

LOTS of great info available at:
fasterdata.es.net/science-dmz

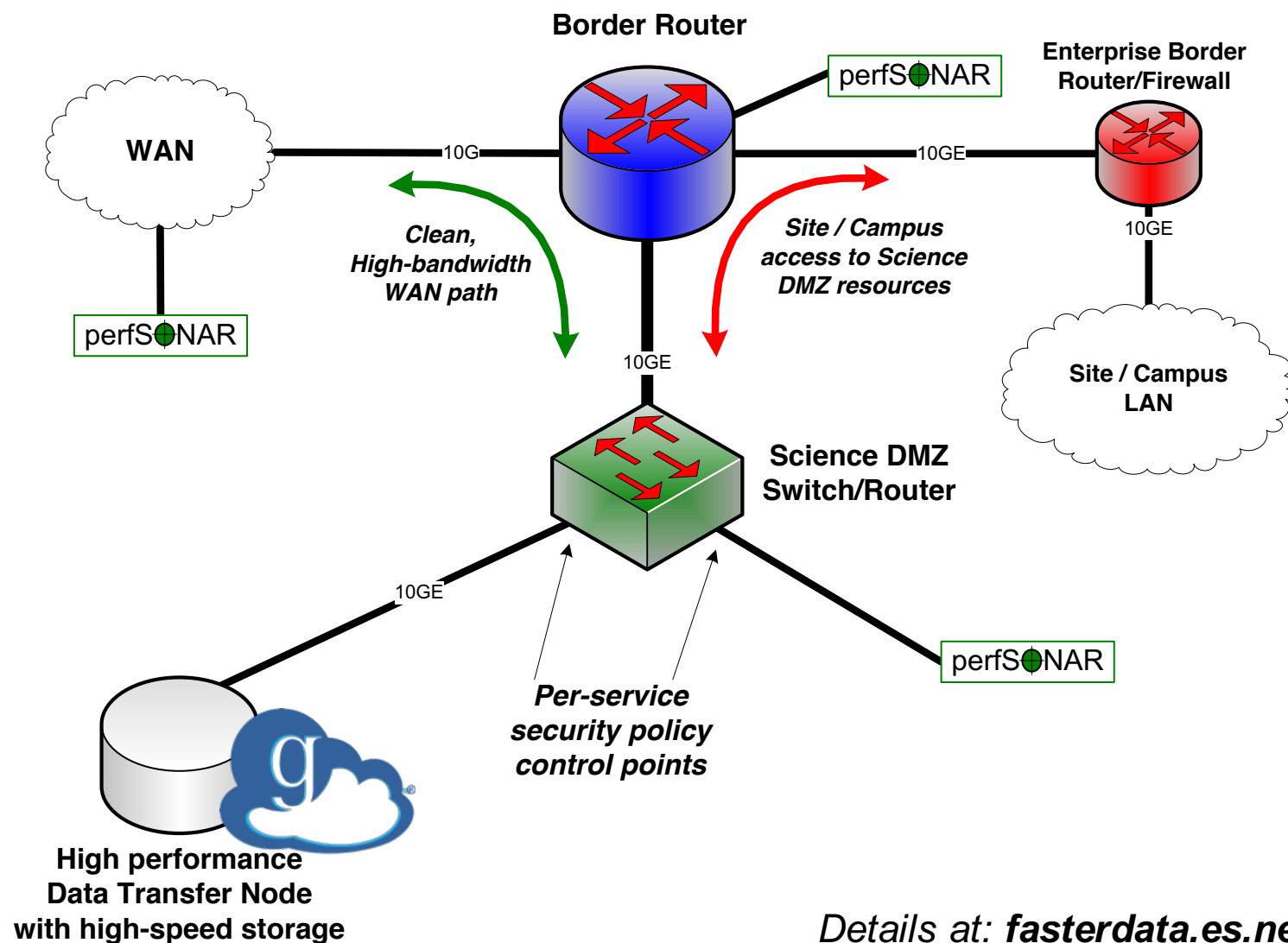


Deployment best practice

Science
DMZ

+

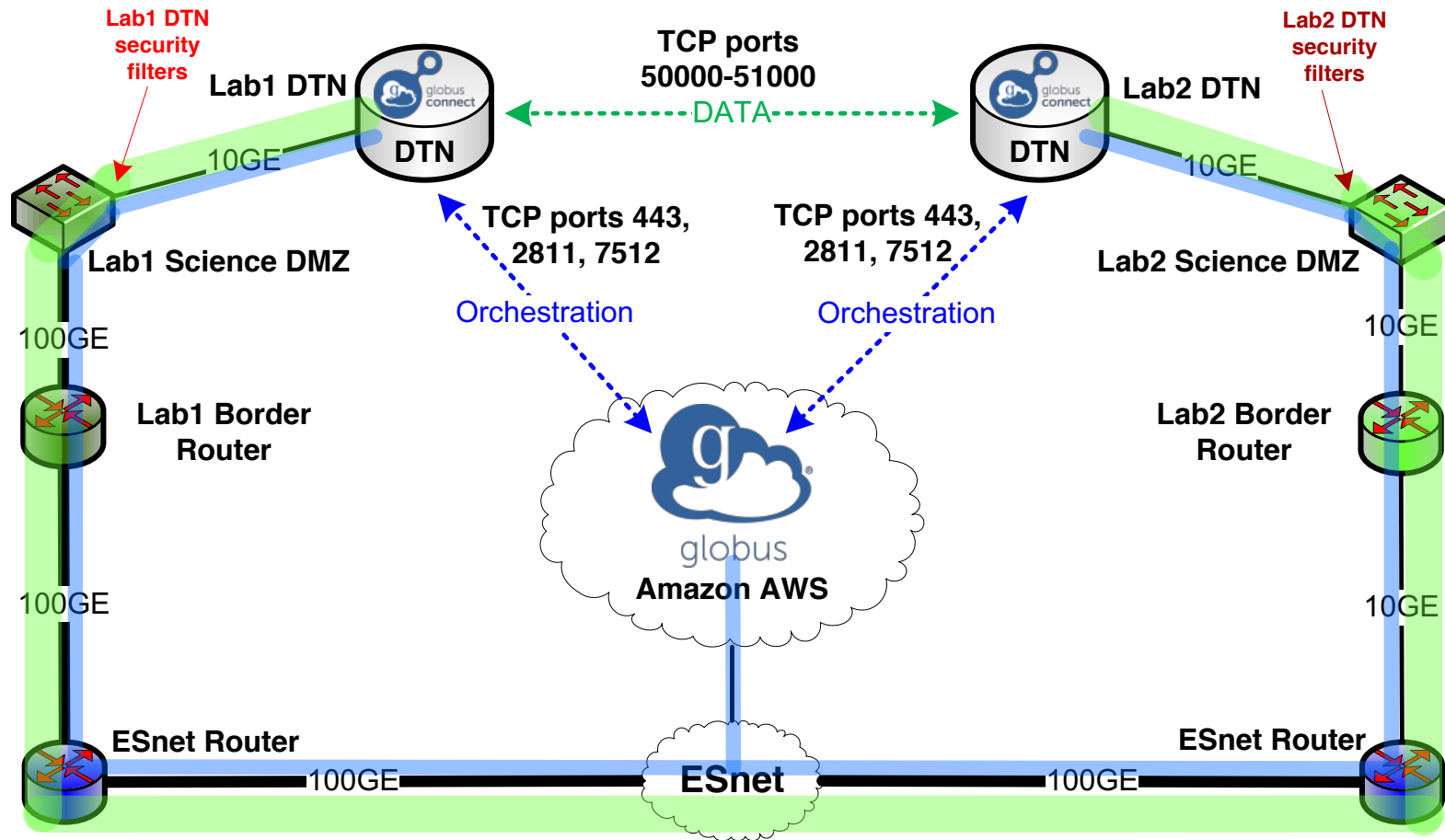
Globus



Details at: fasterdata.es.net



Science DMZ Network paths



Logical data path



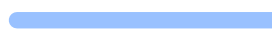
Physical data path



Logical control path



Physical control path





Globus Platform-as-a-Service

XSEDE

Extreme Science and Engineering
Discovery Environment



Globus APIs

Data Publication & Discovery

File Sharing

File Transfer & Replication

Identity, Group, and
Profile Management



Globus Toolkit

Globus Connect





Building bridges to global communities

CISL Research Data Archive

Managed by NCAR's Data Support Section
Data for Atmospheric and Geosciences Research

RDA

Go to Dataset:

- Home
- Find Data
- Ancillary Services
- About/Contact
- Data Citation
- Web Services
- For Staff

Dataset Search

Keyword:

Look For

Recently Added

- The I...
- JRA-5...
- JRA-5...
- JRA-5...
- JRA-5...
- NCEP...
- NCEP...
- JRA-5...
- NCEP...

Find Data

All Datasets | Recently Added/Updated | Browse the RDA

- GCMD Topic:
 - Agriculture • Atmosphere • Biosphere • Climate Indicators • Cryosphere • Hydrosphere • Land Surface • Oceans • Paleoclimate • Solid Earth • Spectral/Engineering • Sun-earth Interactions
- Atmospheric Reanalysis Data:
 - All Reanalysis Datasets • BPRC Arctic System Reanalysis (ASR) • ECMWF 20th Century Reanalysis (ERA-20C) • ECMWF ERA15 Reanalysis (ERA15) • ECMWF ERA40 Reanalysis Project (ERA40) • ECMWF Interim Reanalysis (ERA-I) • JMA Japanese 25-year Reanalysis (JRA25) • JMA Japanese 55-year Reanalysis (JRA55) • NCAR Global Climate Four-Dimensional Data Assimilation Reanalysis (CFDDA) • NCEP Climate Forecast System Reanalysis (CFSR) • NCEP North American Regional Reanalysis (NARR) • NCEP/DOE Reanalysis II (NCEPR2) • NCEP/NCAR Reanalysis Project (NNRP) • NOAA-CIRES 20th Century Reanalysis (20CR)
- Station Observations:
 - Land Surface Air Temperature: Hourly, Monthly

Find Platform Observations datasets

Other Ways to Explore:

- GCMD Topic:
 - Agriculture • Atmosphere • Biosphere • Climate Indicators • Cryosphere • Hydrosphere • Land Surface • Oceans • Paleoclimate • Solid Earth • Spectral/Engineering • Sun-earth Interactions
- Atmospheric Reanalysis Data:
 - All Reanalysis Datasets • BPRC Arctic System Reanalysis (ASR) • ECMWF 20th Century Reanalysis (ERA-20C) • ECMWF ERA15 Reanalysis (ERA15) • ECMWF ERA40 Reanalysis Project (ERA40) • ECMWF Interim Reanalysis (ERA-I)

GLADE Users:

Much of the RDA is directly accessible from CISL's **GLobally Accessible Data Environment**. /glade files can be read directly in place from Yellowstone and Geyser/Caldera. You can find more information under the "Data Access" tab of individual datasets, including detailed lists of /glade files.

Questions

[home page](#)

[September](#)

[September](#)

[August 21,](#)

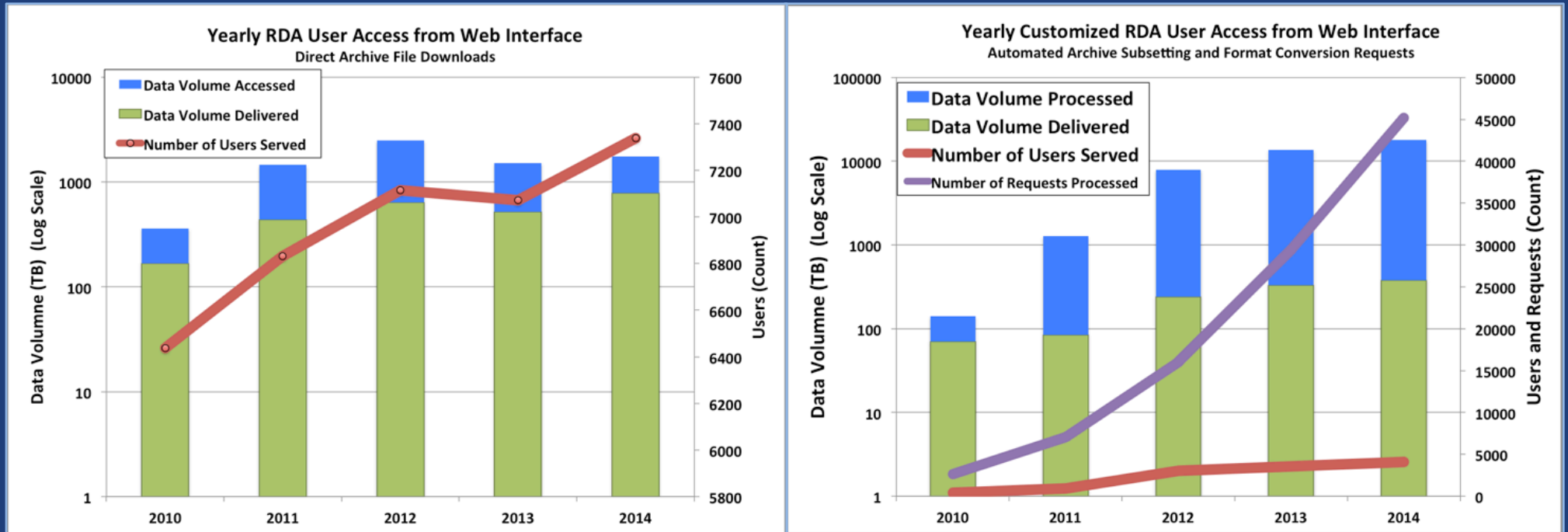


What is the RDA?

- **Free and open access to 600+ datasets for climate and weather research**
- **Worldwide usage**
- **Multiple data access pathways**
 - HTTP (wget, cURL, etc.)
 - OPeNDAP, WCS, WMS
 - Web services (CLI, API)
 - Analysis on HPC systems (NCAR users)



RDA Usage



- **2014**
 - 17+ PB virtual processing
 - Web downloads: 7300 users, 750 TB served
 - 45,000 custom orders, 4000 users, 380 TB served



Globus @ RDA

- **Single shared endpoint**
- **Data copied to subdirectories under endpoint source path**
- **Allow read permission to subdirectories under the shared endpoint**
- **ACLs managed programmatically via Globus CLI**



RDA Alternate Identity login

The screenshot shows the Globus Sign In page. At the top, there is a blue header with the Globus logo and the text 'globus'. To the right of the header are two buttons: 'Log In' and 'Sign Up'. Below the header, the main content area has a 'Sign In' heading on the left and a 'Sign Up with Globus' link on the right. In the center, there are two tabs: 'Using your Globus login.' and 'alternate login'. The 'alternate login' tab is active, and a modal window titled 'Select Identity Provider' is open. This modal contains two columns of identity providers. The first column lists: Globus, Argonne LCF, Argonne MCS & LCRC, BIRN, CLI Transition, EGI, ESG ANL, Exeter, Google, and InCommon / CILogon. The second column lists: LRZ, NCAR RDA, NCSA, NCSA Blue Waters, NERSC, Tuakiri, UChicago CI, UChicago iBi, UK NGS, WestGrid, and XSEDE. The 'NCAR RDA' option in the second column is highlighted with a red rectangular box. To the right of the modal, there is a link that says 'forgot password?'.

globus

Log In Sign Up

Sign In Sign Up with Globus

Using your Globus login. alternate login

Select Identity Provider

Globus	LRZ
Argonne LCF	NCAR RDA
Argonne MCS & LCRC	NCSA
BIRN	NCSA Blue Waters
CLI Transition	NERSC
EGI	Tuakiri
ESG ANL	UChicago CI
Exeter	UChicago iBi
Google	UK NGS
InCommon / CILogon	WestGrid
	XSEDE

forgot password?



RDA Alternate Identity login



NCAR Research Data Archive (RDA) MyProxy Client Authorization

Welcome to the NCAR RDA OAuth for MyProxy Client Authorization Page. The Client below is requesting access to your account. If you approve, please sign in with your RDA email/username and RDA password.

Client Information

Name: Globus Online

URL: <https://www.globusonline.org>

NCAR RDA Email/Username

tcram@ucar.edu



NCAR RDA Password

••••••••••



Sign In

Cancel



Some early Globus supporters

XSEDE

Extreme Science and Engineering
Discovery Environment



PURDUE
UNIVERSITY



Smithsonian

CORNELL
UNIVERSITY

Yale





Enable your campus

- Signup: **globus.org/signup**
- Enable your resource: **globus.org/globus-connect-server**
- Need help? **support.globus.org**
- Subscribe to help make Globus self-sustaining
globus.org/provider-plans
- Follow us: **[@globusonline](https://twitter.com/globusonline)**



Thank you