Installing and Managing Globus on Your Campus Storage Systems

Vas Vasiliadis

vas@uchicago.edu

globus

Useful links/”cheatsheet”:

globusworld.org/**tutorial2016**

Thank you to our sponsors!

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

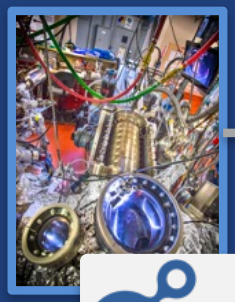
Argonne
NATIONAL LABORATORY

...and Thank YOU!

5 major services	**200PB** transferred	**35 Bn** files processed	**48,000** registered users
13 national labs use Globus	**10,000** active endpoints	**10,000** active users/year	**99.5%** uptime
60+ institutional subscribers	**1 PB** largest single transfer to date	**3 months** longest continuously managed transfer	**150** federated campus identities

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

globus connect

Transfer

1

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

3

Share

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)

globus connect

Publication Repository

- SaaS → Web access; low operational costs
- Use storage system of your choice
- Access using your existing credentials

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

5

Publish

6

8

Personal Computer

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

Discover

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

Analysis store

- Clear staging store
- Data cleansing
- Pre-processing
- ...

Raw Source Data

High durability, low cost store

Bridge to your collaborators

XSEDE
Extreme Science and Engineering Discovery Environment

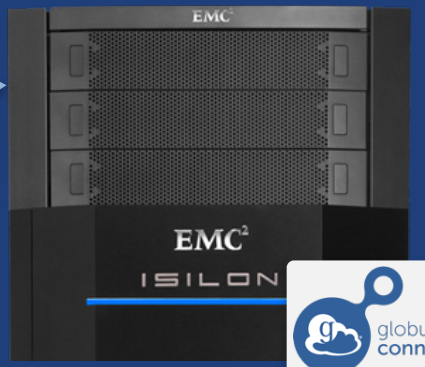
EC2

Jetstream

NERSC

globus connect

Local/Campus stores



globus connect

- Group management
- Share creation
- Permissions configuration

Cloud stores

ceph

openstack

Jetstream

Google Drive

amazon web services S3

globus connect

Bridge to community/public

(Re)format…

- PDF/A
- HDF
- …

Describe…

- Datacite metadata
- Domain metadata
- Provenance info
- ...

Replicated, public repositories

Persistent identifier

- DOI

Multi-step curation workflow

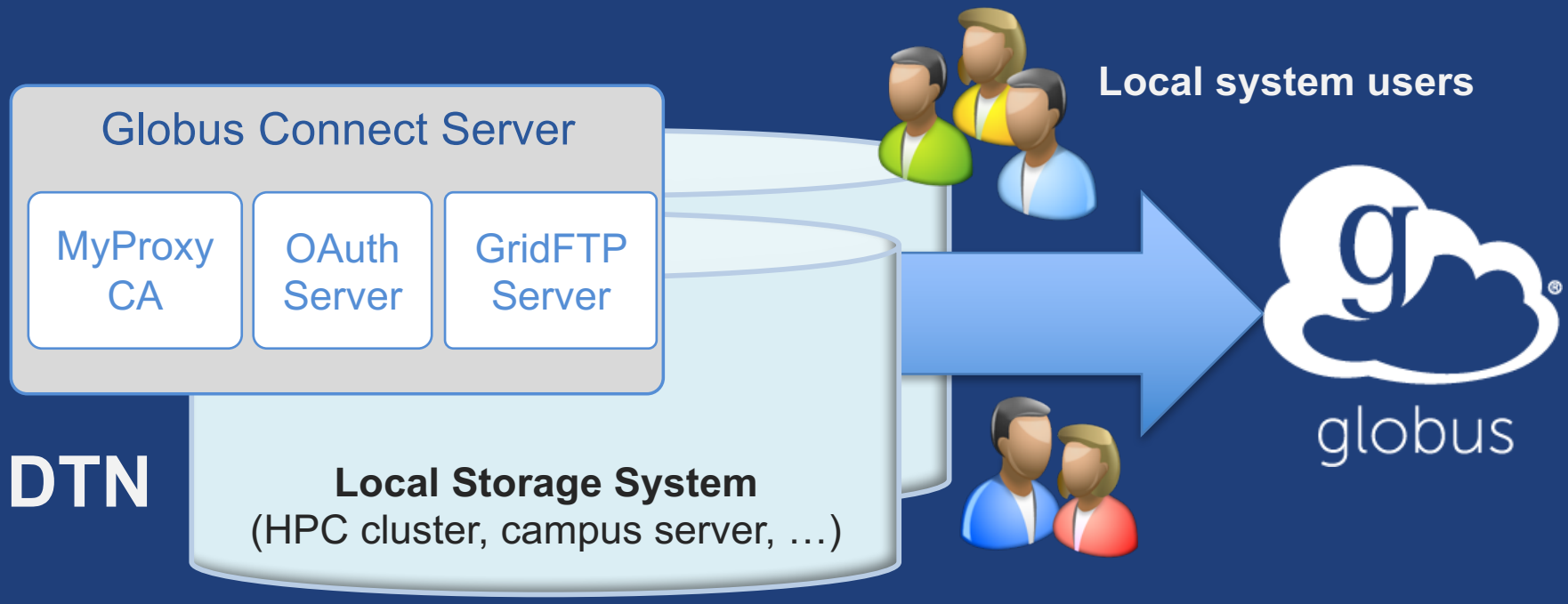
- Review → Update → Resubmit cycle

Demonstration

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**

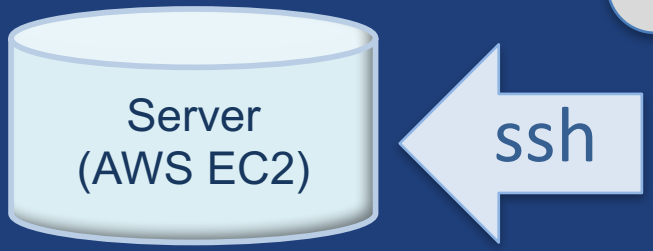
Demonstration

- **Creating a Globus endpoint on your storage system**
- **In this example, storage system = Amazon EC2 server**
- **Akin to what you would do on your DTN**

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`”**

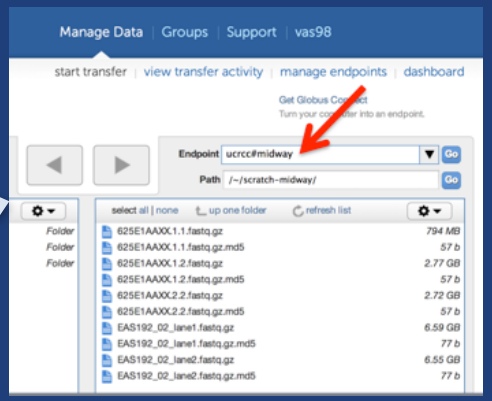
What we are going to do:



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

- **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>`

  (password: **globus2016**)
- **NB: Please** `sudo su` **before continuing**
  - User ‘campusadmin’ has sudo privileges

Step 3: Install Globus Connect Server

Cheatsheet: globusworld.org/tutorial2016

```
$ 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
```

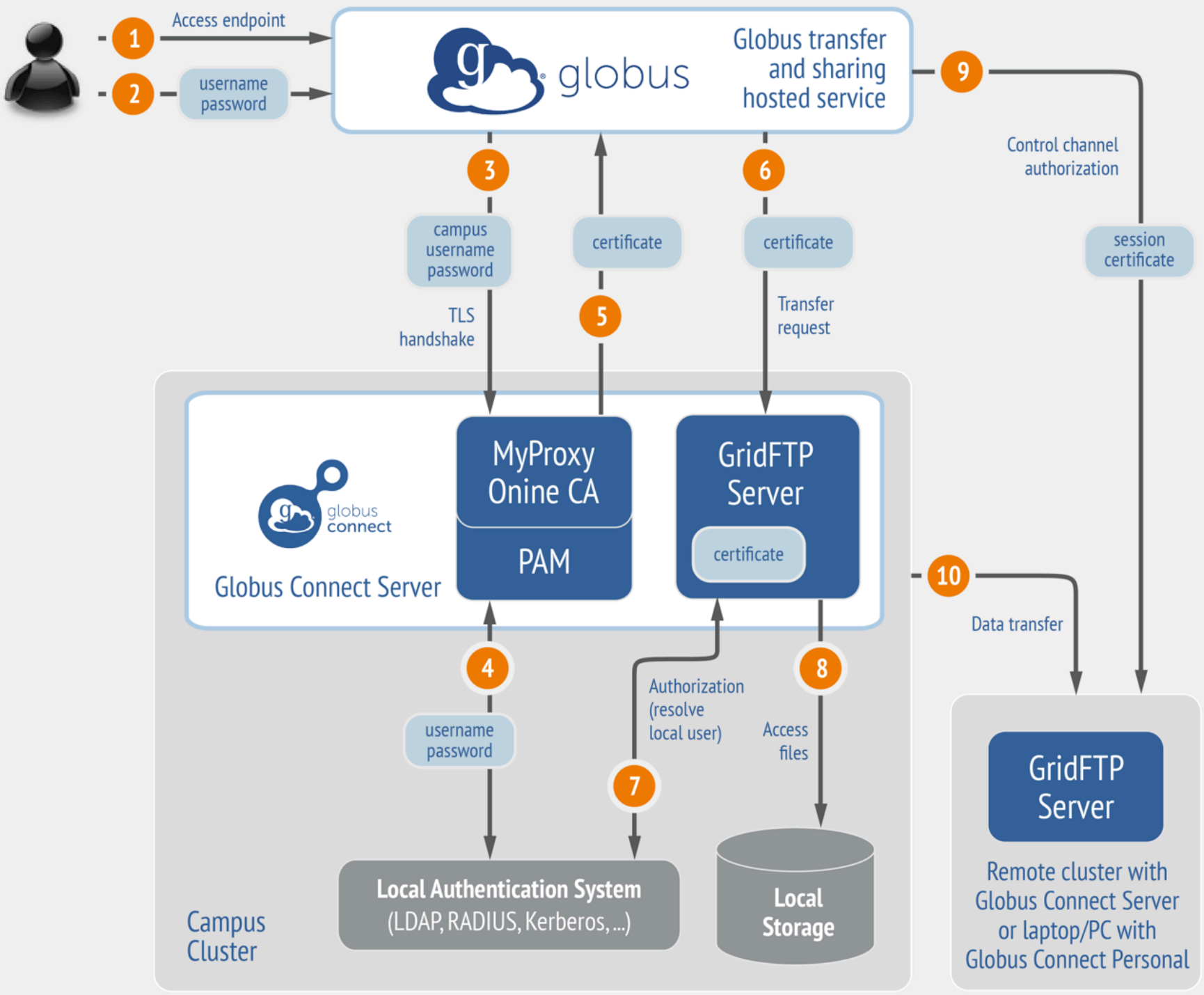
Use your Globus ID username/password when prompted

You have a working Globus endpoint!

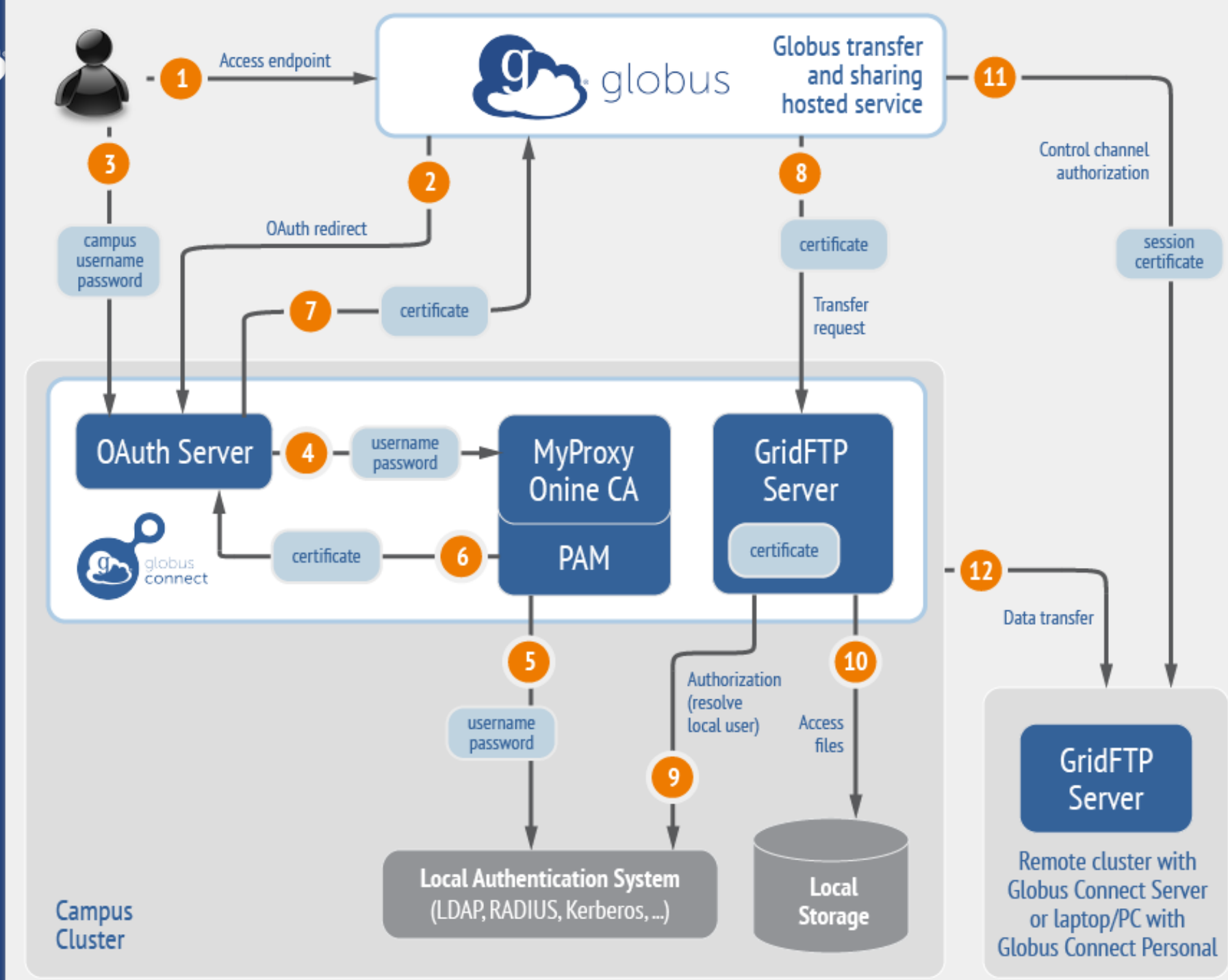
Access the Globus endpoint

- **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**" (pwd: **globus2016**); you should see the user's home directory
- **Transfer files between the ESnet Test endpoints and your endpoint**

Endpoint activation using MyProxy



Endpoint activation using MyProxy OAuth



Ports needed for Globus

- **Inbound: 2811 (control channel)**
- **Inbound: 7512 (MyProxy), 443 (OAuth)**
- **Inbound: 50000-51000 (data channel)**
- **If restricting outbound connections, allow connections from:**
  - 80, 2223 (used during install/config)
  - 50000-51000 (GridFTP data channel)
- **Futures: single-port GridFTP**

Enable your storage system

- Access the service: **globus.org/login**
- Create a personal endpoint: **globus.org/app/endpoints/create-gcp**
- Create a server endpoint: **docs.globus.org/resource-provider-guide**
- Need help? **support@globus.org**
- Subscribe to help us make Globus self-sustaining: **globus.org/subscriptions**
- Follow us: **@globusonline**

Configuring Globus Connect Server

Configuration is a two-step process

- **Configuration options specified in:**
  `/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`**
- **Run `globus-connect-server-setup`**
- **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

- **Set `Sharing = True`**
- **Run `globus-connect-server-setup`**
- **Go to the Transfer Files page**
- **Select the endpoint**
- **Create shared endpoints and grant access to other Globus users***

* 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~/.*`

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`

Access Manager

- **Allow others to manage access to a shared endpoint**
- **Owner of shared endpoint can set role**
- **Assignable to user or group**
- **Common Use Case: Data distribution**

Advanced Configuration

Using MyProxy OAuth server

- **MyProxy without OAuth**
  - Passwords flow via Globus to MyProxy server
  - Globus does not store passwords
  - Still a security concern for many campuses
- **Web-based endpoint activation**
  - Sites run MyProxy OAuth server or use CI Logon
  - Globus gets short-term X.509 credential via MyProxy OAuth protocol

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**

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**

Demonstration:
Command Line Interface (CLI)

Exercise: Globus CLI

1. **Add your SSH key to your Globus ID**
   - Go to: globusid.org/keys
2. `ssh <globusid>@cli.globusonline.org`
3. **Run `help` to see available commands**
4. **Start a transfer and check its `status`**

Using the 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**

Demonstration: Management console

Optimizing transfer performance

Balance: performance - reliability

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

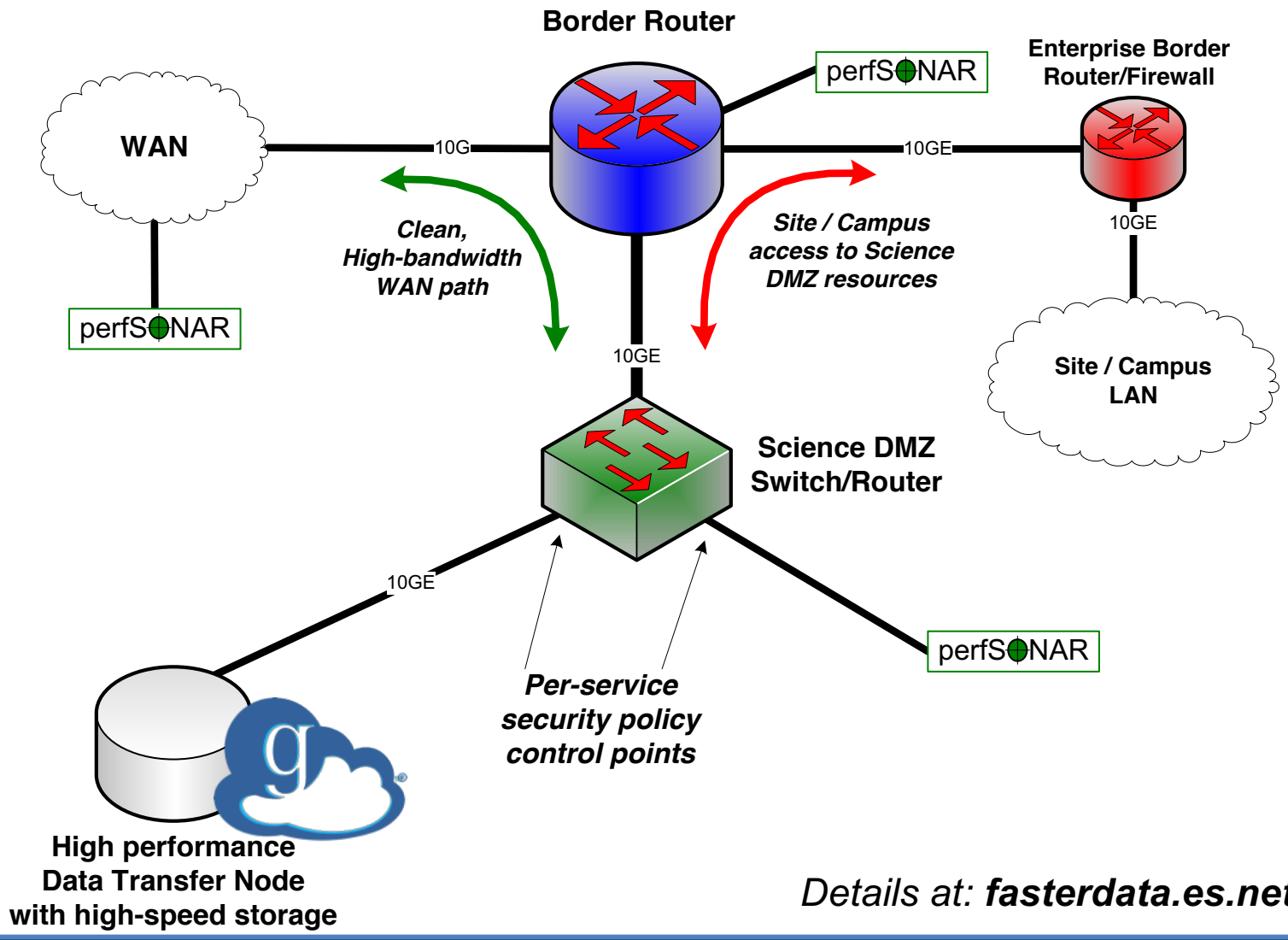
Network Use Parameters

- **Concurrency and parallelism configuration to tune transfers**
- **Maximum and Preferred**
- **Use values set for source and destination to determine parameters for a given transfer**
- **min (max (preferred src, preferred dest), max src, max dest)**

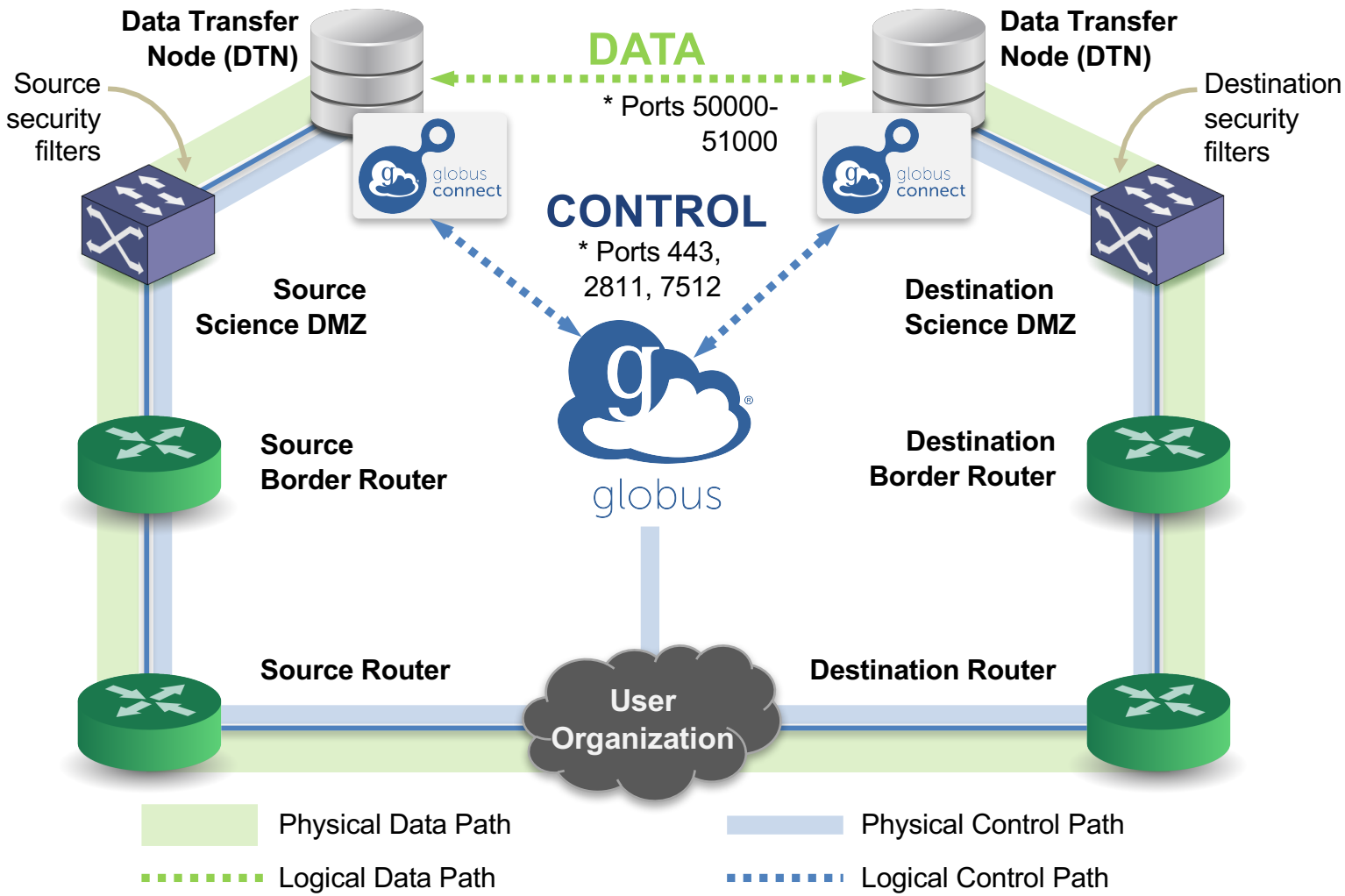
Network paths

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

Best-practice deployment



Network Paths - Illustrative

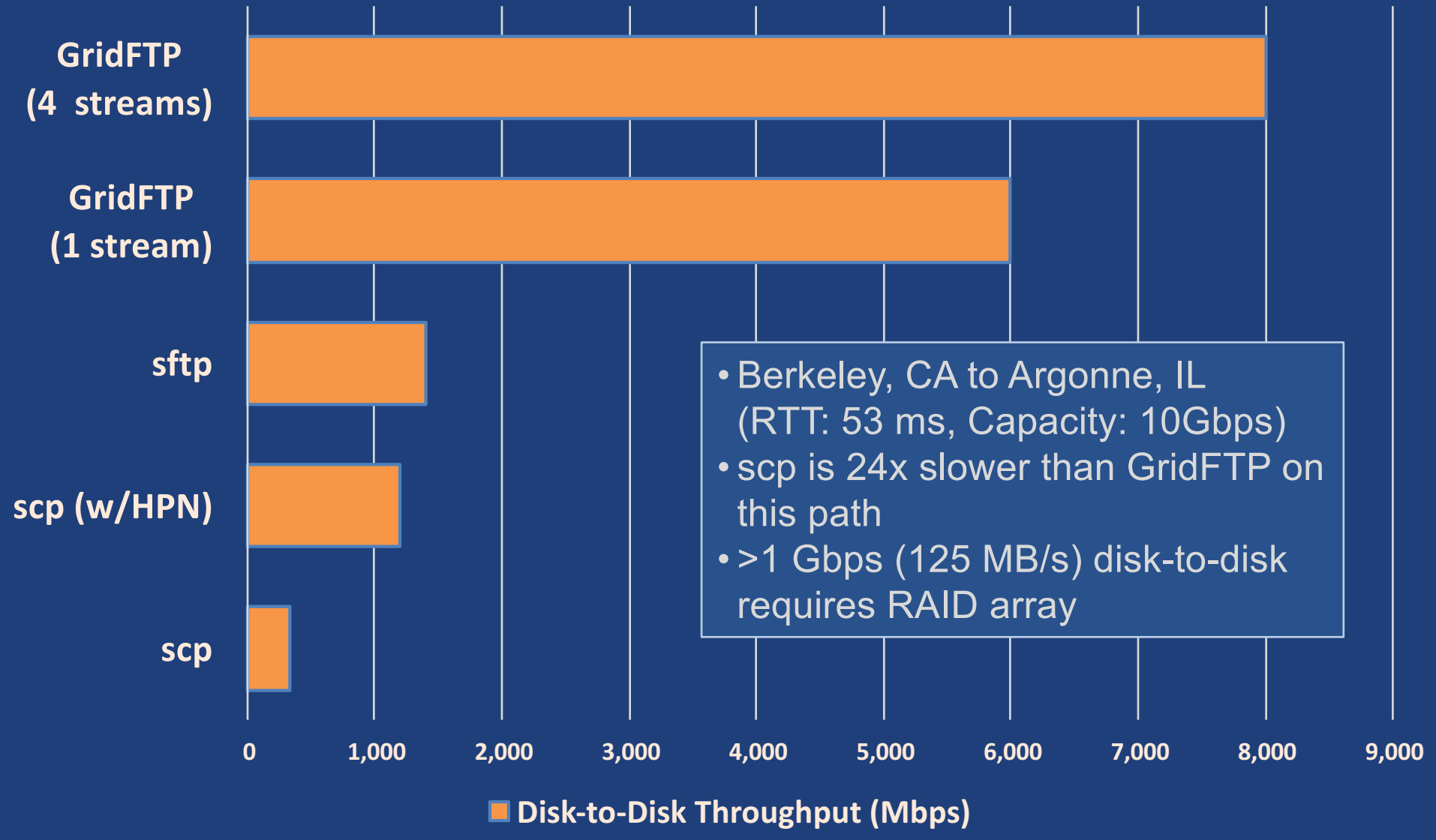


* Please see TCP ports reference: https://docs.globus.org/resource-provider-guide/#open-tcp-ports_section

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)

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**

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 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**

Discussion

Enable your storage system

- Signup: **globus.org/signup**
- Create endpoint: **globus.org/globus-connect-server**
- Need help? **support.globus.org**
- Subscribe to help us make Globus self-sustaining: **globus.org/provider-plans**
- Follow us: **@globusonline**