Registration: bit.ly/ec2ip

1. Select an empty row in the spreadsheet
2. Enter your name and email address
3. Make a note of the IP address displayed

Slides and useful links:
globusworld.org/tutorials
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 (…)

MIRA
Bridge to instruments

Pre-processed Data

High durability, low cost store

Raw Source Data

Analysis store

Amazon Glacier
Bridge to collaborators

External Campus Storage

Public/PrivateCloud stores

EC2

XSEDE

Jetstream

NERSC

Harvard

Cornell University

red cloud

ceph

openstack

Google Drive

amazon web services

S3
Bridge to community/public

Project Repositories, Replication Stores

Public Repositories

EC2

XSEDE

Jetstream

NERSC
Globus SaaS: Research data lifecycle

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

2. Globus transfers files reliably, securely.

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

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

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

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

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

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

- Use a Web browser
- Access any storage
- Use an existing identity
Conceputal architecture: Hybrid SaaS

**Subscriber Control Domain**
- Subscriber owned and administered storage system
- "client" software

**Globus Control Domain**
- Single, globally accessible multi-tenant service

**Source Endpoint**
- Globus connect

**Destination Endpoint**
- No data relay or staging via Globus

**DATA Channel**
- Source Endpoint to Destination Endpoint

**CONTROL Channel**
- Subscriber Control Domain
- Globus Control Domain
Conceptual architecture: Sharing

External User Control Domain

DATA Channel

Managed Endpoint

Subscriber managed filesystem permissions

CONTROL Channel

Globus Connect

Globus Control Domain

Shared Endpoint

Globus managed "overlay" permissions

Subscriber Control Domain

Managed Endpoint

Globus managed "overlay" permissions
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
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
Moving data between your laptop and another system
Exercise: Log in & transfer files

- Go to: www.globus.org/login
- Select your institution from the list and click “Continue” (or log in with a Globus ID)
- Authenticate with your institution’s identity system
- Install Globus Connect Personal
- Move some data between an ESnet test endpoint and your laptop
Sharing Data
Exercise: Share files

• Join the “Tutorial Users” groups
  – Go to “Groups”, search for “tutorial”
  – Select group from list, click “Join Group”

• Create a shared endpoint on your laptop

• Grant your neighbor permissions on your shared endpoint

• 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

Globus service

Web

CLI

Rest API

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

(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

Commands:

bookmarks  Manage Endpoint Bookmarks
config     Modify, view, and manage your Globus CLI config.
Globus as PaaS

Integrate file transfer and sharing capabilities into scientific web apps, portals, gateways, etc.

Use existing institutional ID systems in external web applications.
Data Distribution: ARM Climate Research Facility
Data Distribution: APS - DMagic

dmagic.readthedocs.io

Courtesy of Francesco De Carlo, Argonne National Laboratory (2016)
Analysis App: Wellcome Sanger

Sanger Imputation Service

This is a free genotype imputation and phasing service provided by the Wellcome Trust Sanger Institute. You can upload GWAS data in VCF or 23andMe format and receive imputed and phased genomes back. Click here to learn more and follow us on Twitter.

Before you start

Be sure to read through the instructions. You will need to set up a free account with Globus and have Globus Connect running at your institute or on your computer to transfer files to and from the service.

Ready to start?

If you are ready to upload your data, please fill in the details below to register an imputation and/or phasing job. If you need more information, see the about page.

- Full name
- Organisation
- Email address

What is this?
Globus user identity

News

@sangerimpute

11/05/2016
Thanks to EAGLE, we can now return phased data. The HRC panel has been updated to r1.1 to fix a known issue. See ChangeLog for more details.

15/02/2016
Globus API changed, please see updated instructions.

17/12/2015
New status page and reworked internals. See ChangeLog.

09/11/2015
Pipeline updated to add some features requested by users. See ChangeLog.
National Resource Access

Jetstream Web App would like to:

- Access all Jetstream resources.

By clicking “Allow”, you allow Jetstream Web information and services. You can rescind this permission at any time.

Compute Canada has partnered with Globus to offer this high performance file transfer service.

Calculate Canada s'est associé à Globus pour vous offrir ce service de transfert de fichier à haute performance.

Log in to use Compute Canada Globus Web App

Use your existing organizational login

- e.g. university, national lab, facility, project, Google or Globus ID

Your Globus username and password used prior to February 13, 2016 is now Globus ID.

- WestGrid

Continue

Didn’t find your organization? Then use Globus ID to sign up.
Globus PaaS developer resources

Python SDK

Sample Application

Docs.globus.org/api

Github.com/globus

Jupyter Notebook

Requirements
- You need to be in the tutorial users group for sharing: https://www.globus.org/app/groups/50b6a29c-63ac-11e4-8062-22000ab8755
- Installed Globus Python SDK

Configuration
First you will need to configure the client with an OAuth2 access token. For the purpose of this tutorial, you can use this website. Click the "Jupyter Notebook" option and copy the resulting text below, or click on "Globus CLI" and
Thank you to our sponsors…
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…

- 360 PB transferred
- 62 billion tasks processed
- 73,000 registered users
- 10,000 active endpoints
- 500 100TB+ users
- 14,000 active users
- 3 months longest running managed transfer
- 100TB+ transferred
- 350+ federated identities
- 1 PB largest single transfer to date
- 5,000 active shared endpoints
- 99.5% uptime
...and THANK YOU to our subscribers
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

Stanford University – February 8, 2018
Get IP address: bit.ly/ec2ip

‘campusadmin’ - sudo privileges
‘researcher’ - regular user
Globus Connect Server

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

docs.globus.org/globus-connect-server-installation-guide/
Globus Connect Server

Non-POSIX Connectors

POSIX-compliant Connector

MyProxy CA
OAuth Server
GridFTP Server

DTN

Local Storage System
(HPC cluster, NAS, ...)

Local system users
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_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
Exercise: Make your endpoint visible

• **Edit endpoint attributes**
  – Change the name to something useful, e.g. `<your_name> EC2 Endpoint`
  – For the “Visible To” attribute select “Public - Visible to all users”

• **Find your neighbor’s endpoint**
  – You can access it too 😊
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?**
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

1. Access endpoint
2. Username password
3. Campus username password
4. Username password
5. TLS handshake
6. Certificate
7. Username password
8. Authorization (resolve local user) 
9. Control channel authorization
10. Data transfer

DON’T LEAVE IT LIKE THIS!
Yes, please do this!
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
  \[
  \min \left( \max(\text{preferred src}, \text{preferred dest}), \right.
  \left. \max \text{ src}, \right. \left. \max \text{ dest} \right)
  \]
• 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: ESnet Testing**

- **GridFTP (4 streams)**
- **GridFTP (1 stream)**
- **sftp**
- **scp (w/HPN)**
- **scp**

*Disk-to-Disk Throughput (Mbps)*

- Berkeley, CA to Argonne, IL (RTT: 53 ms, Capacity: 10Gbps)
- scp is 24x slower than GridFTP on this path
- >1 Gbps (125 MB/s) disk-to-disk requires RAID array

Source: ESnet (2016)
Deployment Scenarios
Best practice network configuration

DATA
* Ports 50000-51000

CONTROL
* Ports 443, 2811, 7512

Source Security filters
Destination Security filters

Source Router
Destination Router

Source Science DMZ
Destination Science DMZ

Source Border Router
Destination Border Router

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

On-premises Data Transfer Node (DTN)

Data Storage Interface (DSI) for POSIX-compliant filesystems

Non-POSIX DSI
- Google Drive
- Amazon S3 (native)
- Spectra BlackPearl
- Ceph S3 RadosGW
- HPSS

Cloud-hosted DTN

AWS EBS Volume

AWS S3 Bucket

Non-POSIX DSI
- Google Drive
- Amazon S3 (native)
- Spectra BlackPearl
- Ceph S3 RadosGW
- HPSS
Multi-endpoint configuration

Data Transfer Node

~/
~/.scratch

POSIX DSI

ext*
XFS
ZFS
Multi-endpoint configuration

- Data Transfer Node
- POSIX DSI
- ext*, XFS, ZFS
- ~/scratch
- ~/projects
- GPFS Lustre
- ~/projects

Transfer Files

[Image of a Globus interface]
Multi-endpoint configuration

Data Transfer Node

- POSIX DSI
- Spectra BlackPearl DSI

~/.projects
~/.archive
~/.scratch

../archive

ext*, XFS, ZFS

GPFS Lustre

~/projects
~/archive
~/scratch
Multi-endpoint configuration

Data Transfer Node

~/
~/.scratch

POSIX DSI

~/projects

Spectra BlackPearl DSI

~/archive

Amazon S3 DSI

~/vault

ext* XFS ZFS

GPFS Lustre

Amazon S3 Bucket

Transfer Files

Transfer Files

Transfer Files
Deploying a premium connector gateway

- **Object Store / Archive Connector**
- **Globus endpoint** (e.g. Amazon S3 “gateway”)
- **Other Globus endpoints**
- **Data**
- **Control**

Diagram illustrating the deployment of a premium connector gateway with connections to Globus Share, Object Store/Archive, and server endpoints.
Network paths

• Separate control and data interfaces
• "DataInterface =" option in globus-connect-server-conf
• Common scenario: route data flows over Science DMZ link
Dual-homed DTN – high speed data path

GridFTP Server

Science DMZ

Data Transfer Node

Data Transfer Node

Internet2 path

GridFTP Data Channel

GridFTP Server

GridFTP Control Channel

if0

if1

globus
Dual-homed DTN – private network data path

Science DMZ

Data Transfer Node

Firewall

GridFTP Server

GridFTP Control Channel

if0

if1

LAN/Intranet path

GridFTP Data Channel

GridFTP Server
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: ensure use of FIPS capable OpenSSL libraries on DTN
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
  – `-id` on separate server, e.g. non-DTN nodes
  – `-web` on either `-id` server or separate server for OAuth interface
Globus Connect Server Deployment

Science DMZ (ACL limited)
Port 2811 accepts inbound connections from Globus

Data Transfer Node

GridFTP Server

MyProxy CA

ext* XFS ZFS

OAuth Server

Port 2811

Firewall

Science DMZ

Microsoft AD

OpenID Connect
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

• **Install Globus Connect Server on all servers**

• **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**

• **Copy the configuration file to all servers**
  – /etc/globus-connect-server.conf

• **Run globus-connect-server-setup on the server running the –id component**

• **Run globus-connect-server-setup on all other servers**

• **Repeat steps 2-5 as necessary to update configurations**
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
Globus Network Manager
For environments with super duper special network constraints…
(a.k.a. ”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**
Future directions
Motivations for Globus Connect Server v5

• Facilitate automation of installation and upgrades
• Allow scale out deployment
  – Across DTNs
  – Across multiple file system connectors
• Reduce number of ports required
• Streamline user experience with use of Globus sharing
• Enhance user registration of credentials for cloud storage connectors
• Prepare foundation for next set of enhanced capabilities
New features with Globus Connect Server v5

- **Collection model**
- **HTTPS access to storage**
- **Security improvements**
  - OAuth2 in GridFTP (no more X.509 user certificates or Myproxy!)
  - OpenID Connect identity provider
  - Credential expiration LoA policies
  - User credential management (e.g., for Google Drive, S3, Kerberos)
- **Kerberos protected file systems**
- **Directory listing with path expressions**
Collections: The evolution of endpoints
Collection properties

- Set of blobs (files), hierarchically named (folders)
- Rooted at a unique DNS name
- URL referenceable files, folders
- Accessible and manageable via:
  - HTTPS: client/server file access
  - GridFTP: async bulk transfer
  - REST API: advanced operations

- OAuth2 authentication and authorization via Globus Auth
- Collection-specific access policies
- Data is stored on a storage system, which determines storage policies such as durability and availability
- File change events
Installation & configuration enhancements for v5

• Setup with any identity (GlobusID not required)
• Automatable installation and configuration
• Configuration API, CLI, GUI
• Scale-out deployment without shared file system
• Backup / restore configuration to / from the cloud
• Multiple storage systems simultaneously
• Single port GridFTP (no ephemeral ports)
• Distributed as Docker containers
Streamlined data sharing with v5

• **Remove friction of sharing**
  – Guest collections where possible, e.g., Google Drive
  – Hybrid collections: Mapped access to home & project folders, else guest access

• **Enhanced sharing permissions**
  – permission expiration
  – permissions on files (not just folders)
  – sharing via URL possession

• **Storage connectors: share from anywhere**
New capabilities built on collections and v5

• **Data search (early release available now by request)**
  – With access control
  – Schema agnostic
  – Custom indexes domain specific

• **Event driven actions for automation**
  – Replication of data (across storage tiers)
  – Metadata extraction and ingest to search
  – Run analysis pipelines
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
Support resources

- Customer engagement team
- Globus documentation: docs.globus.org
- Helpdesk and issue escalation: support@globus.org
- Globus professional services team
  - Assist with portal/gateway/app architecture and design
  - Develop custom applications that leverage the Globus platform
  - Advise on customized deployment and integration scenarios
Open Discussion