Introduction to Globus:

Research Data Management Software at the ALCF

Rick Wagner

rick@globus.org
rpwagner@uchicago.edu
rwagner@anl.gov

Research data management today

Index?

How do we...

...move?

...share?

...discover?

...reproduce?

Globus delivers…

Fast and reliable ~~big~~ data transfer, sharing, and platform services…

…directly from your own storage systems…

...via software-as-a-service using existing identities with the overarching goal of...

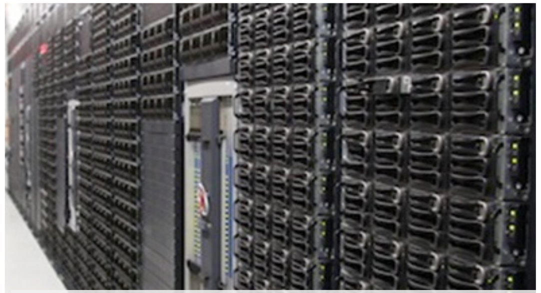
…unifying access to data across tiers

Welcome to
compute
canada

ACCESS SERVICES AND EXPERTS

DISCOVER COMPUTE CANADA

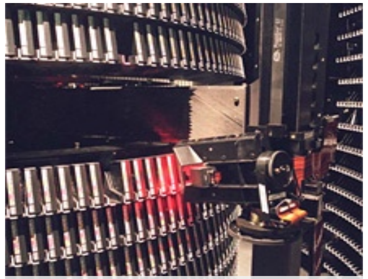
National Resources


Research Computing HPC

Personal Resources

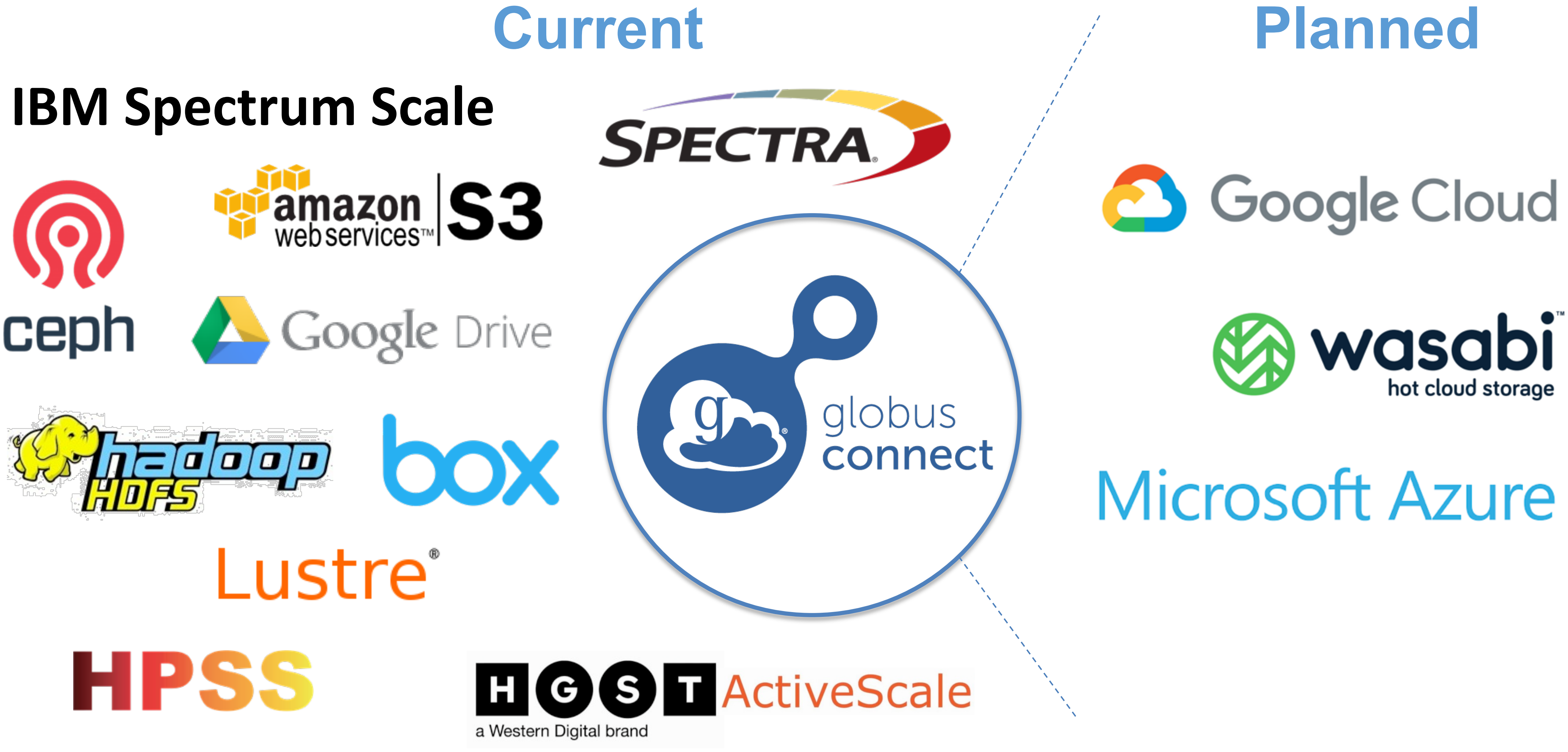
Public Cloud

Desktop Workstations


Mass Storage

Instruments

Storage Connectors - globus.org/connectors



Current

IBM Spectrum Scale

SPECTRA

ceph

amazon web services™ S3

Google Drive

hadoop HDFS

box

Lustre®

HPSS

HGST ActiveScale

a Western Digital brand

globus connect

Planned

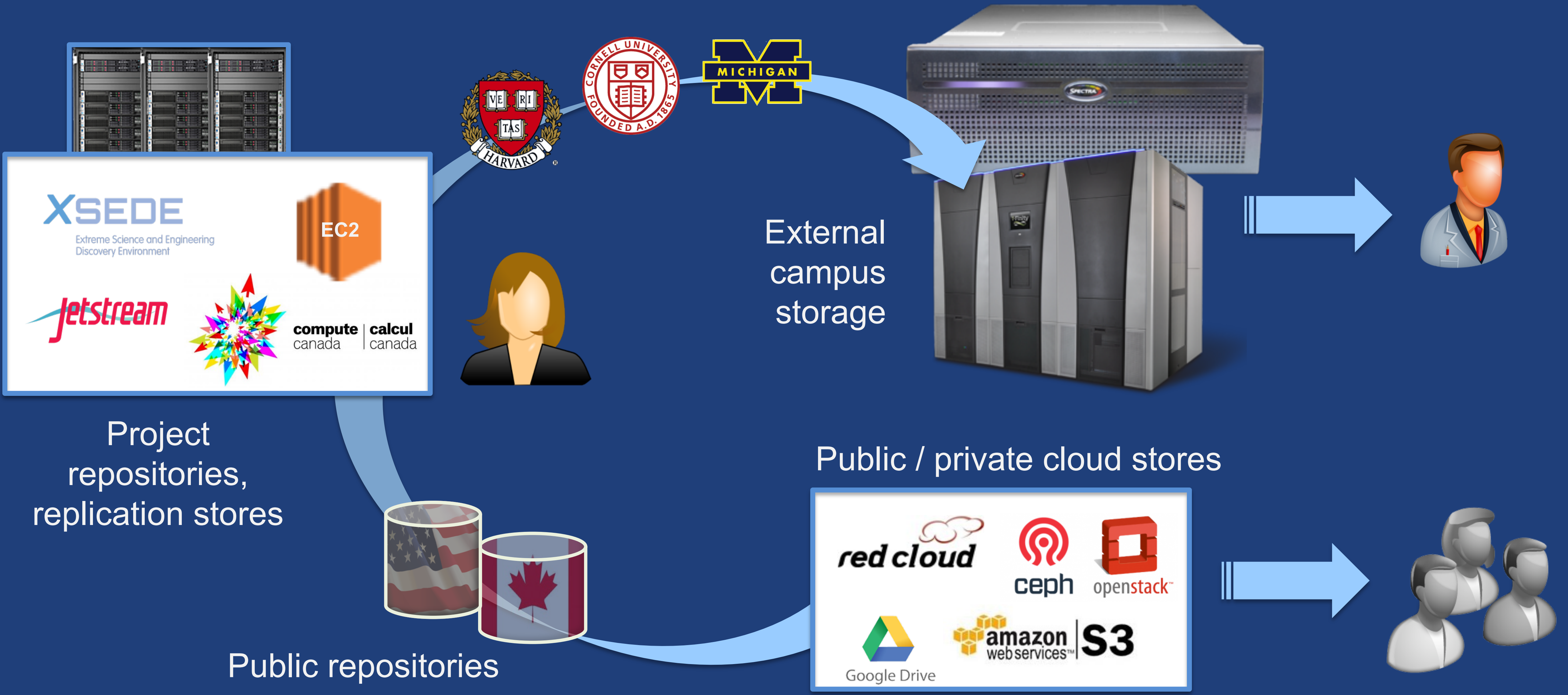
Google Cloud

wasabi™

hot cloud storage

Microsoft Azure

Share with collaborators/community



External campus storage

Project repositories, replication stores

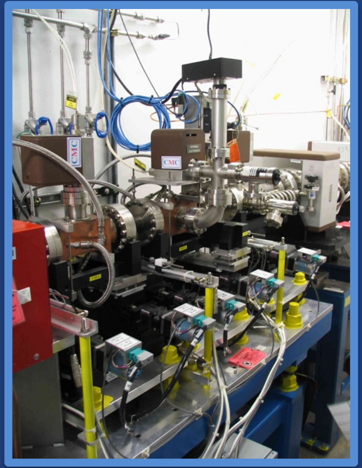
Public repositories

Public / private cloud stores

VE RI TAS
HARVARD
CORNELL UNIVERSITY FOUNDED A.D. 1865
MICHIGAN
XSEDE
Extreme Science and Engineering Discovery Environment
EC2
Jetstream
compute canada | calcul canada
red cloud
ceph
openstack
Google Drive
amazon web services
S3

Manage data from instruments

Next-Gen Sequencer



Advanced Light Source

Cryo-EM

MRI

Light Sheet Microscope

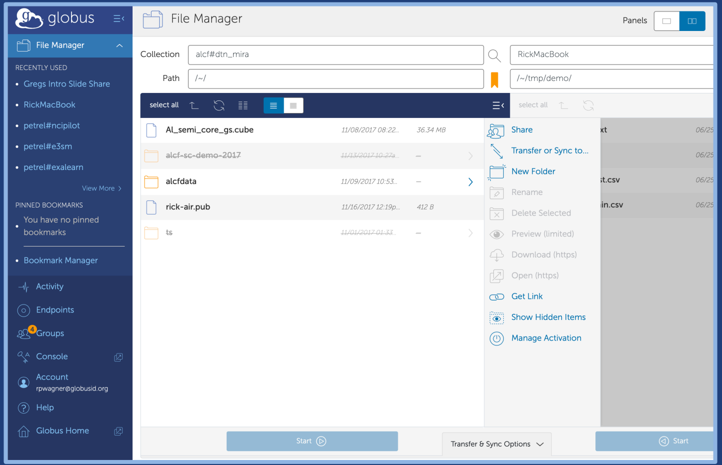
Analysis store

High-durability, low-cost store

Remote visualization

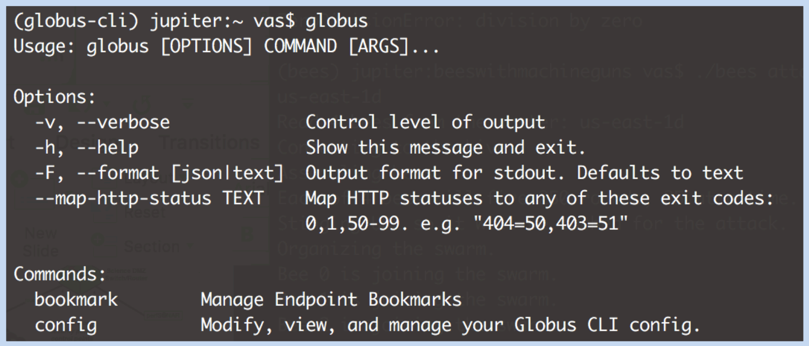
Personal system

Use(r)-appropriate interfaces



Web

Globus service



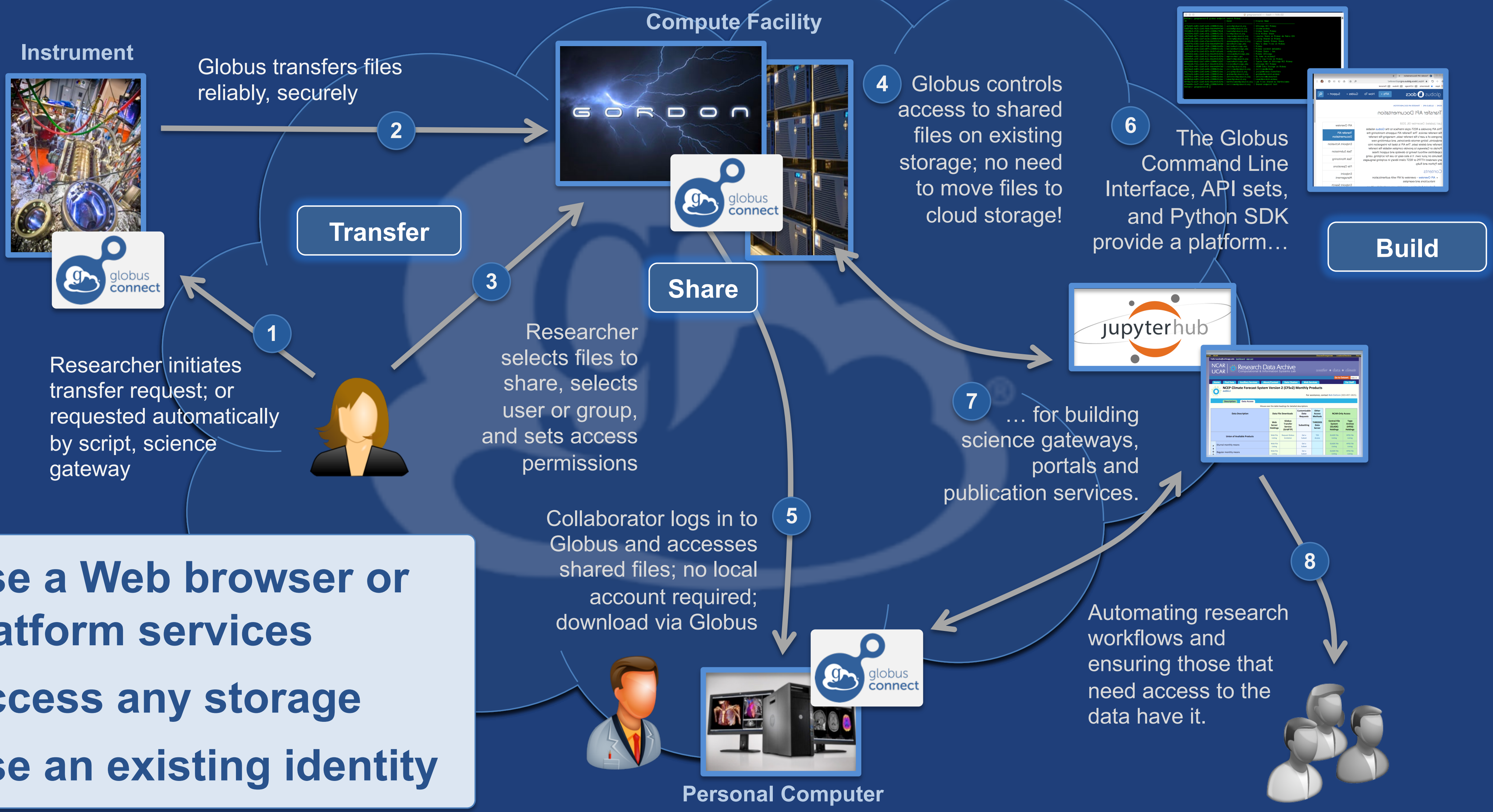
```
(globus-cli) jupiter:~ vas$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

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 SaaS / PaaS: 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

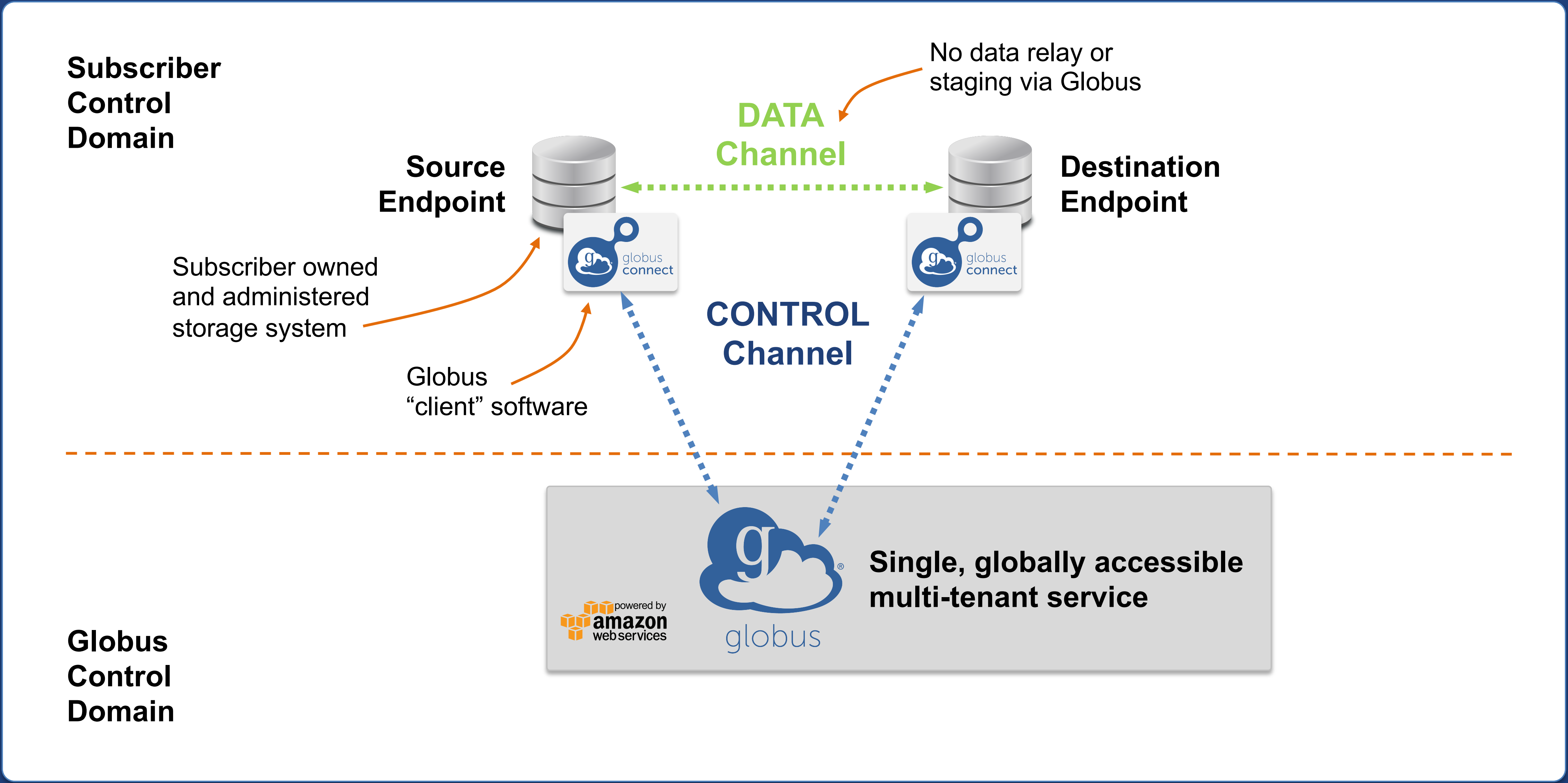
6 The Globus Command Line Interface, API sets, and Python SDK provide a platform…

7 … for building science gateways, portals and publication services.

8 Automating research workflows and ensuring those that need access to the data have it.

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

Conceptual architecture: Hybrid SaaS



Subscriber Control Domain

No data relay or staging via Globus

DATA Channel

Source Endpoint

Destination Endpoint

Subscriber owned and administered storage system

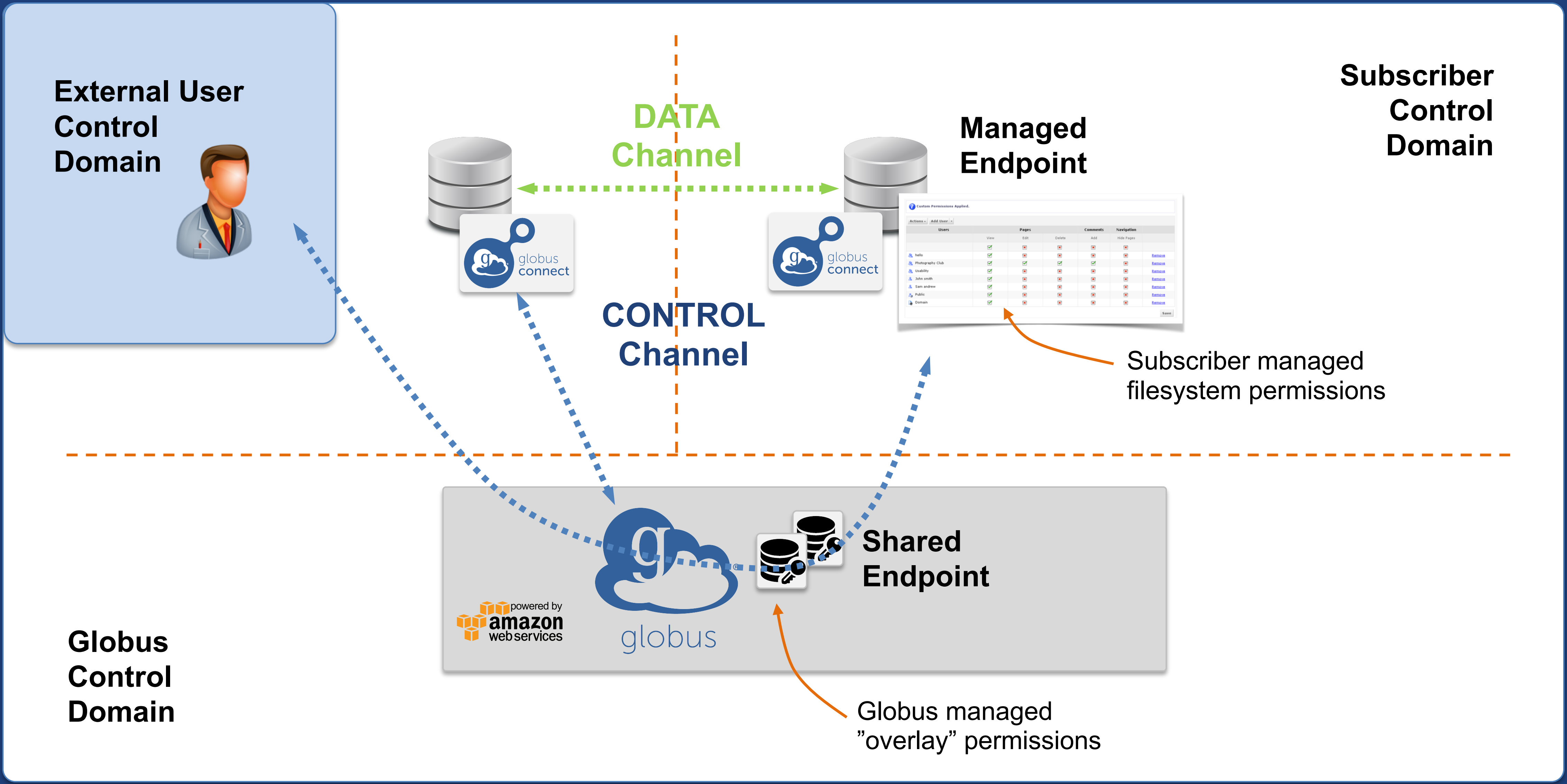
Globus "client" software

CONTROL Channel

Single, globally accessible multi-tenant service

Globus Control Domain

Conceptual architecture: Sharing



External User Control Domain

DATA Channel

CONTROL Channel

Managed Endpoint

Subscriber Control Domain

Custom Permissions Applied.

Actions ▾ | Add User ▾

Users	Pages			Comments	Navigation	
	View	Edit	Delete	Add	Hide Pages	
	✓	☒	☒	☒	☒	
hello	✓	☒	☒	☒	☒	Remove
Photography Club	✓	✓	✓	✓	☒	Remove
Usability	✓	☒	☒	☒	☒	Remove
John smith	✓	☒	☒	☒	☒	Remove
Sam andrew	✓	☒	☒	☒	☒	Remove
Public	✓	☒	☒	☒	☒	Remove
Domain	✓	☒	☒	☒	☒	Remove

Save

Subscriber managed filesystem permissions

Shared Endpoint

Globus managed ”overlay” permissions

Globus Control Domain

…makes your storage system a Globus endpoint

Endpoints (Collections)

- **Storage abstraction**
  - All transfers happen between two endpoints
  - Globus Connect instantiates endpoints
- **Collection ~= Endpoint**
- **Test / Demo Endpoints**
  - Globus Tutorial Endpoint 1
  - Globus Tutorial Endpoint 2
  - ESnet Test Endpoints
    - Contain file samples of various sizes
- **Globus Connect Personal**
  - Now your laptop is an endpoint
  - https://www.globus.org/globus-connect-personal

Globus Connect Personal

- Installers do not require admin access
- Zero configuration; auto updating
- Handles NATs
- Installs in seconds – easy to delete

The Globus Web App - Accounts

- **A Globus Account is**
  - A Primary Identity
  - Possible Linked Identities
- **Linking Identities**
- **Managing Identities**
- **Consents**

Demonstration

Identities

File Transfer

File Sharing

Activity Monitoring

- **Recent / History / Filter**
- **Drilling Down**
  - File transfer statistics
  - Overview
  - Event Log
  - Cancelling an active task

Groups

- **What can they be used for?**
  - Sharing: Access permissions for more than one person
  - Roles: Endpoint management and monitoring
- **Groups**
  - Creating groups and setting the visibility
  - Members (invitations), Subgroups, Settings
  - Settings
    - Policies / Membership Fields / Terms & Conditions
  - Roles
    - Giving others authority over your groups

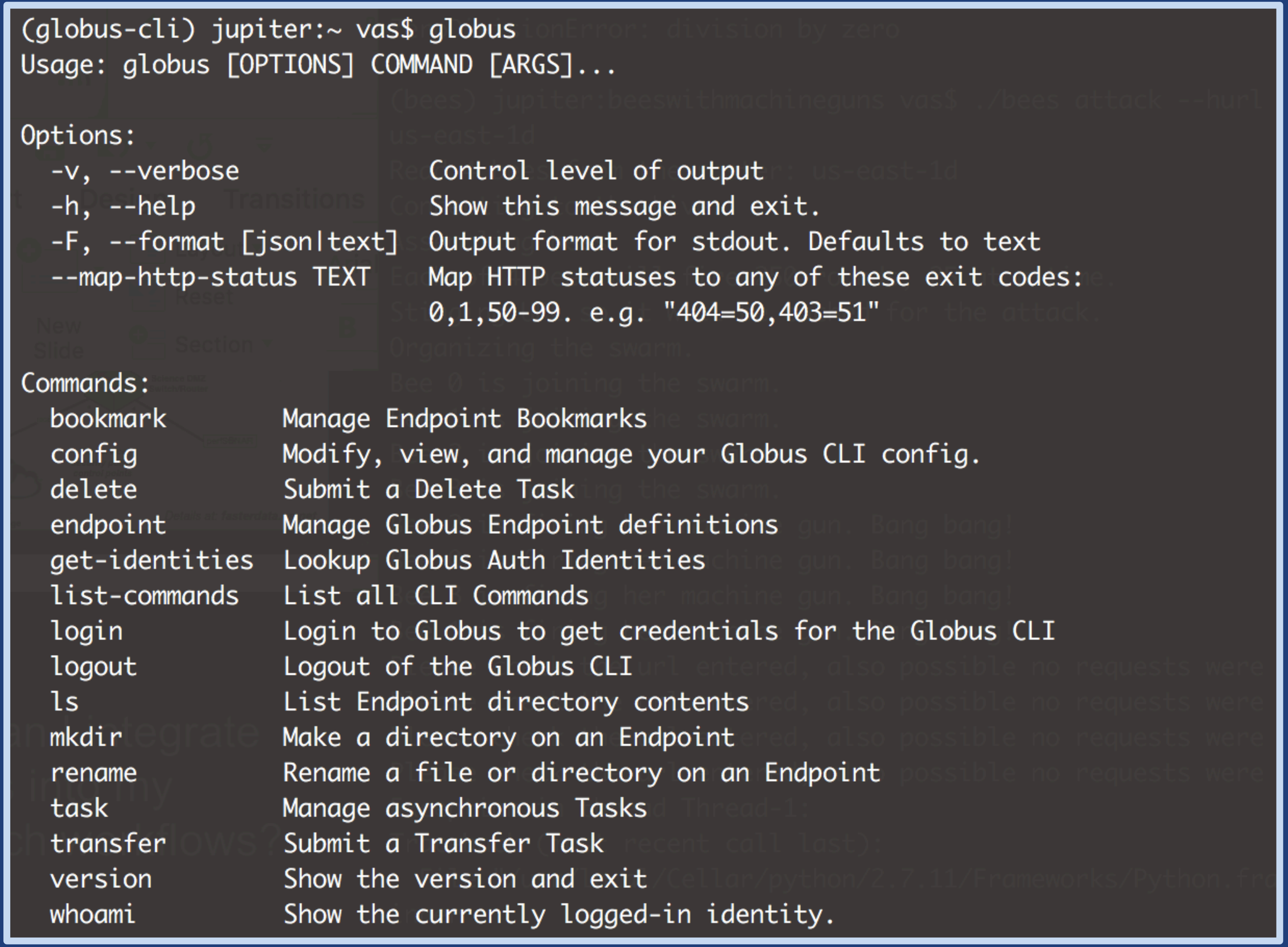
Endpoint Sharing and Roles

- **Sharing**
  - Select the directory and create the “share”
  - A ”share” is another type of endpoint
  - Share with: Users / Groups / All Globus Users
- **Roles**
  - Giving others (or groups of others) control or monitoring rights for your endpoints

Bookmarks

- Just like browser bookmarks – frequently used, or maybe not used frequently enough!
- Creating a bookmark
- Using a bookmark
- Sorting and Filtering
- Editing and Deleting

Globus Command Line Interface



```
(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.
  delete          Submit a Delete Task
  endpoint        Manage Globus Endpoint definitions
  get-identities  Lookup Globus Auth Identities
  list-commands   List all CLI Commands
  login           Login to Globus to get credentials for the Globus CLI
  logout          Logout of the Globus CLI
  ls              List Endpoint directory contents
  mkdir           Make a directory on an Endpoint
  rename          Rename a file or directory on an Endpoint
  task            Manage asynchronous Tasks
  transfer        Submit a Transfer Task
  version         Show the version and exit
  whoami          Show the currently logged-in identity.
```

Open source, uses Python SDK

docs.globus.org/cli

github.com/globus/globus-cli

The Globus CLI

- **Installation**
  - docs.globus.org/cli/installation
  - Prerequisites
- **Logging On (remember the consents?)**
  - globus login / logout
- **Getting help / list of commands**
  - globus –help
  - globus list-commands
- **Doing something**
  - It all about the UUIDs
  - Don’t forget the file paths!

The Globus CLI – Let’s do a few things…

- **Find endpoints**
  - globus endpoint search Midway
  - globus endpoint search ESNet
  - globus endpoint search --filter-scope=recently-used
- **Find endpoint contents**
  - globus ls af7bda53-6d04-11e5-ba46-22000b92c6ec
  - globus ls af7bda53-6d04-11e5-ba46-22000b92c6ec:RMACC2018
- **Transfer a file**
  - From ESnet Read-Only Test DTN at CERN to Midway
  - Note the specific paths
  - globus transfer d8eb36b6-6d04-11e5-ba46-22000b92c6ec:/~/data1/1M.dat af7bda53-6d04-11e5-ba46-22000b92c6ec:/~/1M.dat
- **Transfer a directory**
  - From Globus Tutorial Endpoint 2 to Midway (create directory and contents)
  - globus transfer --recursive ddb59af0-6d04-11e5-ba46-22000b92c6ec:/~/sync-demo af7bda53-6d04-11e5-ba46-22000b92c6ec:/~/syncDemo
- **https://docs.globus.org/cli/examples/**

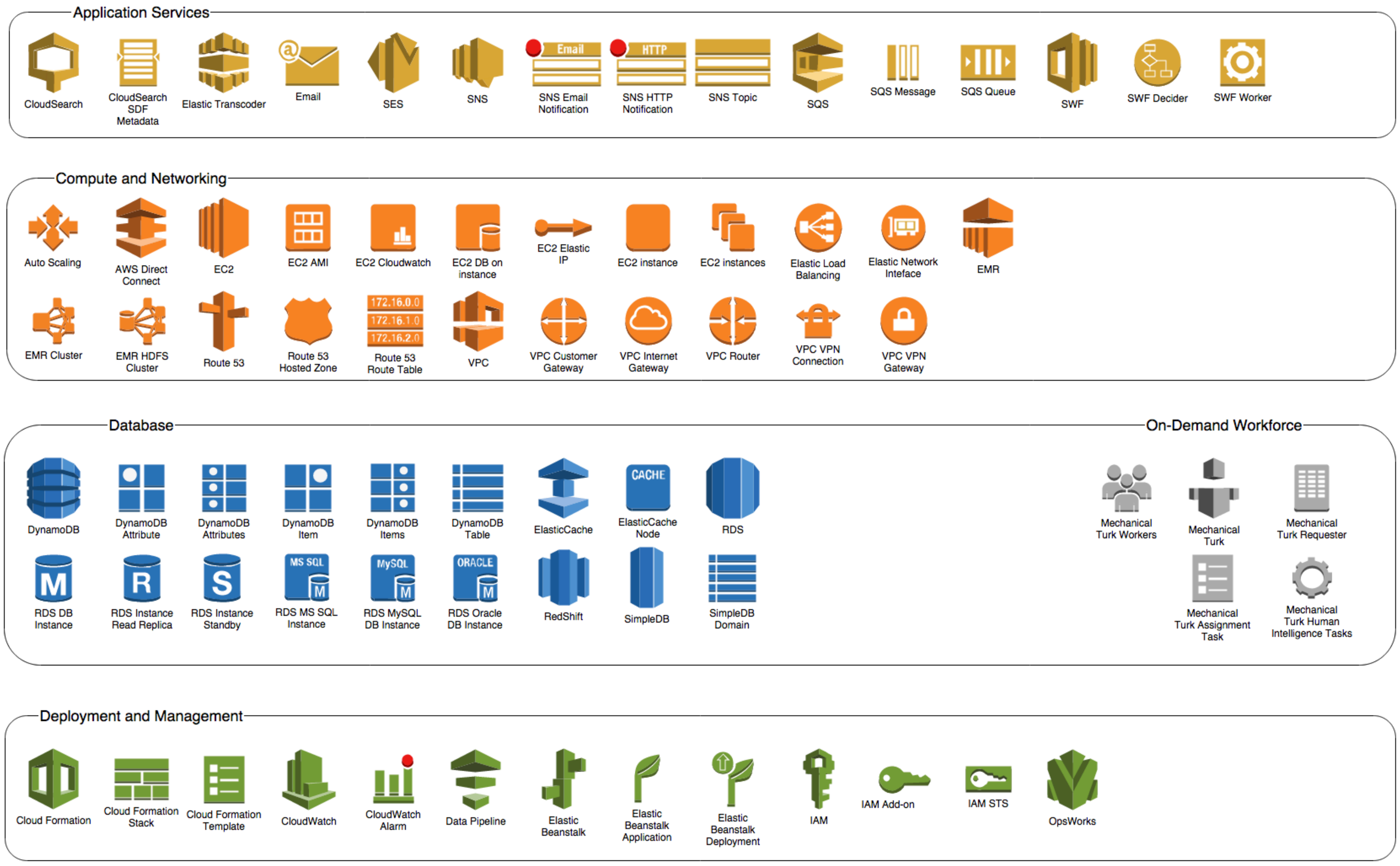
Globus CLI

- **Easy install and updates**
- **It’s a native application distributed by Globus**
  - https://docs.globus.org/cli/
  - https://github.com/globus/globus-cli
- **Command *globus login* gets access tokens and refresh tokens**
  - Stores the token locally (~/.globus.cfg)
  - The CLI ”acts as” the logged in user
- **All interactions with the service use the tokens**
  - Tokens for Globus Auth and Transfer services
- **Command *globus logout* deletes those**
- **https://docs.globus.org/cli/examples/**
- **https://github.com/globus/automation-examples**

Demonstration

Globus CLI

Industry software builds on platform services



Application Services

CloudSearch, CloudSearch SDF Metadata, Elastic Transcoder, Email, SES, SNS, SNS Email Notification, SNS HTTP Notification, SNS Topic, SQS, SQS Message, SQS Queue, SWF, SWF Decider, SWF Worker

Compute and Networking

Auto Scaling, AWS Direct Connect, EC2, EC2 AMI, EC2 Cloudwatch, EC2 DB on instance, EC2 Elastic IP, EC2 instance, EC2 instances, Elastic Load Balancing, Elastic Network Inteface, EMR

EMR Cluster, EMR HDFS Cluster, Route 53, Route 53 Hosted Zone, Route 53 Route Table, VPC, VPC Customer Gateway, VPC Internet Gateway, VPC Router, VPC VPN Connection, VPC VPN Gateway

Database

DynamoDB, DynamoDB Attribute, DynamoDB Attributes, DynamoDB Item, DynamoDB Items, DynamoDB Table, ElastiCache, ElastiCache Node, RDS

RDS DB Instance, RDS Instance Read Replica, RDS Instance Standby, RDS MS SQL Instance, RDS MySQL DB Instance, RDS Oracle DB Instance, RedShift, SimpleDB, SimpleDB Domain

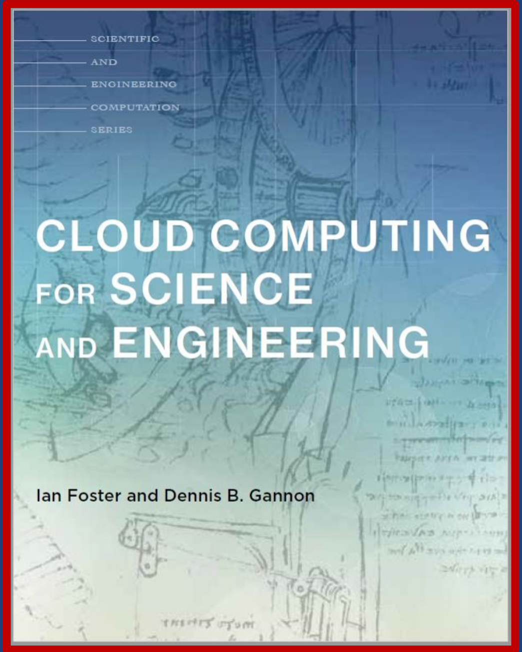
On-Demand Workforce

Mechanical Turk Workers, Mechanical Turk, Mechanical Turk Requester, Mechanical Turk Assignment Task, Mechanical Turk Human Intelligence Tasks

Deployment and Management

Cloud Formation, Cloud Formation Stack, Cloud Formation Template, CloudWatch, CloudWatch Alarm, Data Pipeline, Elastic Beanstalk, Elastic Beanstalk Application, Elastic Beanstalk Deployment, IAM, IAM Add-on, IAM STS, OpsWorks

cloud4scieng.org



SCIENTIFIC AND ENGINEERING COMPUTATION SERIES

CLOUD COMPUTING FOR SCIENCE AND ENGINEERING

Ian Foster and Dennis B. Gannon

Globus delivers… with applications and as a platform…

Fast and reliable data transfer, sharing, and file management…

…directly from your own storage systems…

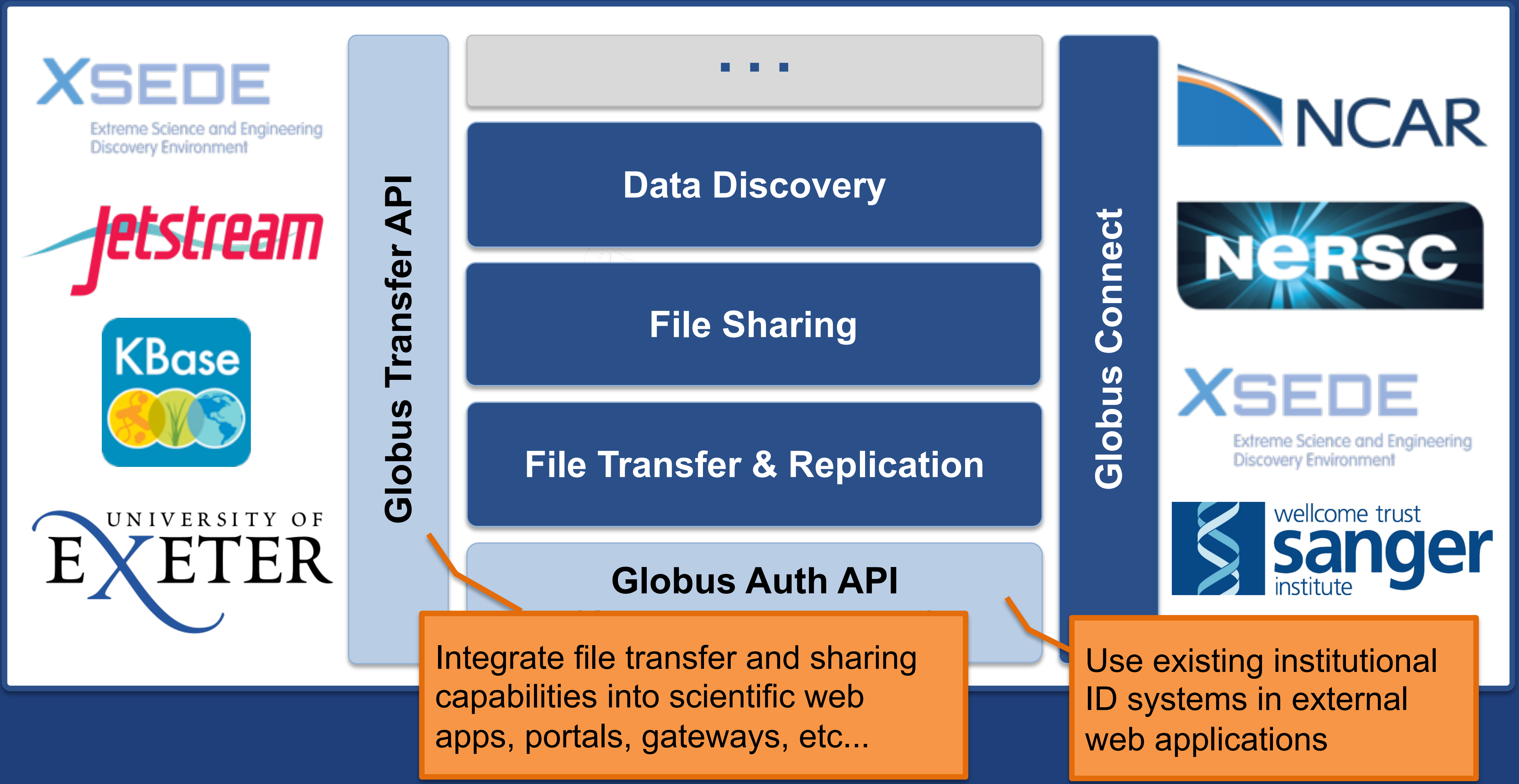
...via software-as-a-service using existing identities.

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.

Globus Platform-as-a-Service

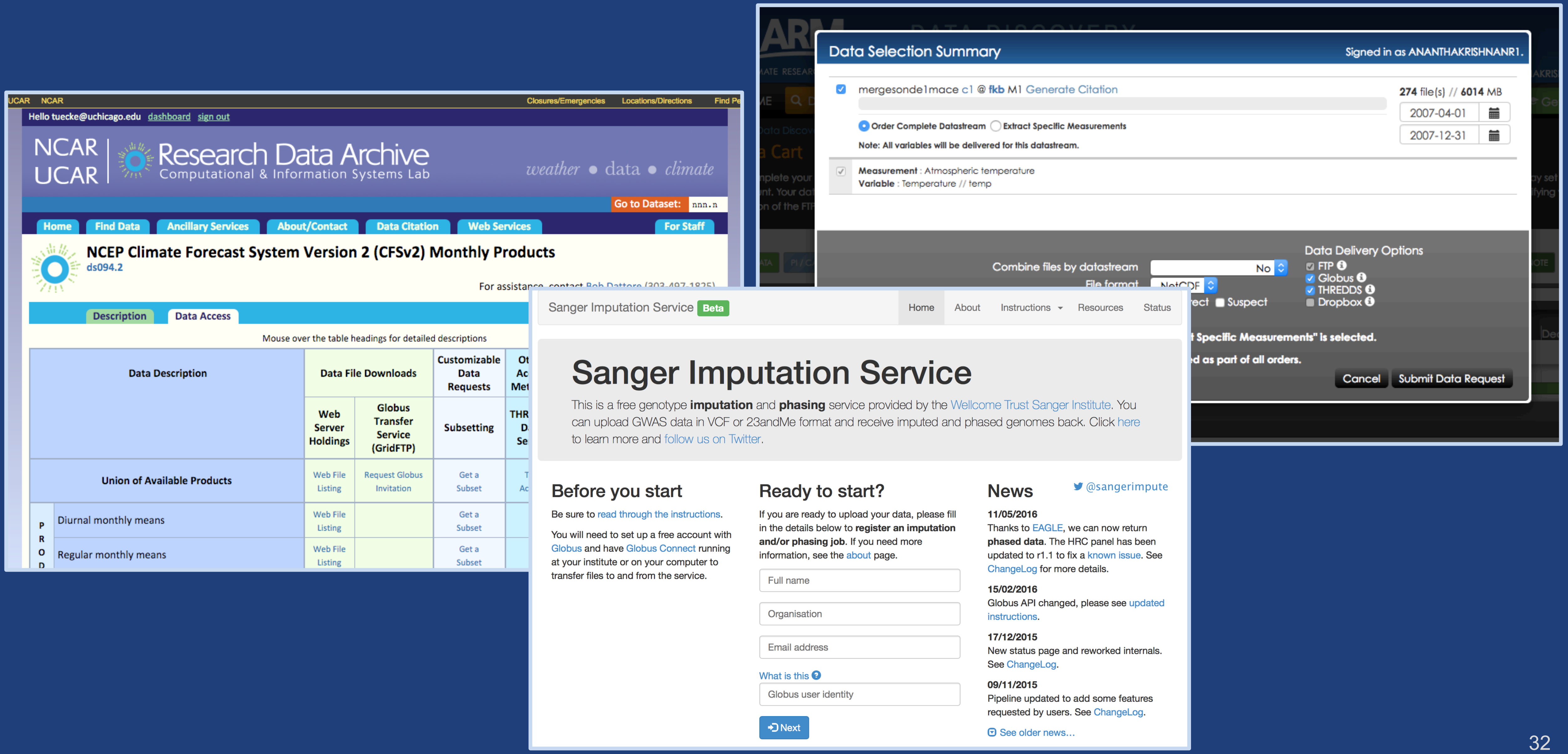


XSEDE
Extreme Science and Engineering Discovery Environment
Jetstream
KBase
UNIVERSITY OF EXETER
Globus Transfer API
. . .
Data Discovery
File Sharing
File Transfer & Replication
Globus Auth API
Globus Connect
NCAR
NERSC
XSEDE
Extreme Science and Engineering Discovery Environment
wellcome trust sanger institute

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

Use existing institutional ID systems in external web applications

Example web apps that leverage Globus



UCAR NCAR Closures/Emergencies Locations/Directions Find Pe

Hello tuecke@uchicago.edu dashboard sign out

NCAR UCAR | Research Data Archive
Computational & Information Systems Lab

weather • data • climate

Go to Dataset: nnn.n

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

NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products

ds094.2

For assistance, contact Bob Dattore (303-497-1825)

Description | Data Access

Mouse over the table headings for detailed descriptions

Data Description		Data File Downloads		Customizable Data Requests	Ot Ac Met
		Web Server Holdings	Globus Transfer Service (GridFTP)	Subsetting	THR D Se
Union of Available Products		Web File Listing	Request Globus Invitation	Get a Subset	T Ac
P R O D	Diurnal monthly means	Web File Listing		Get a Subset	
	Regular monthly means	Web File Listing		Get a Subset	

ARM DATA DISCOVERY

Data Selection Summary

Signed in as ANANTHAKRISHNANR1.

☑ mergesonde1mace c1 @ fkb M1 Generate Citation

274 file(s) // 6014 MB

2007-04-01

2007-12-31

● Order Complete Datastream ○ Extract Specific Measurements

Note: All variables will be delivered for this datastream.

☑ **Measurement** : Atmospheric temperature
Variable : Temperature // temp

Combine files by datastream: No

File format: NetCDF

ect □ Suspect

Data Delivery Options
- ☑ FTP
- ☑ Globus
- ☑ THREDDS
- □ Dropbox

t Specific Measurements" is selected.

d as part of all orders.

Cancel | Submit Data Request

Sanger Imputation Service Beta — Home | About | Instructions | Resources | Status

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

Next

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.

See older news...

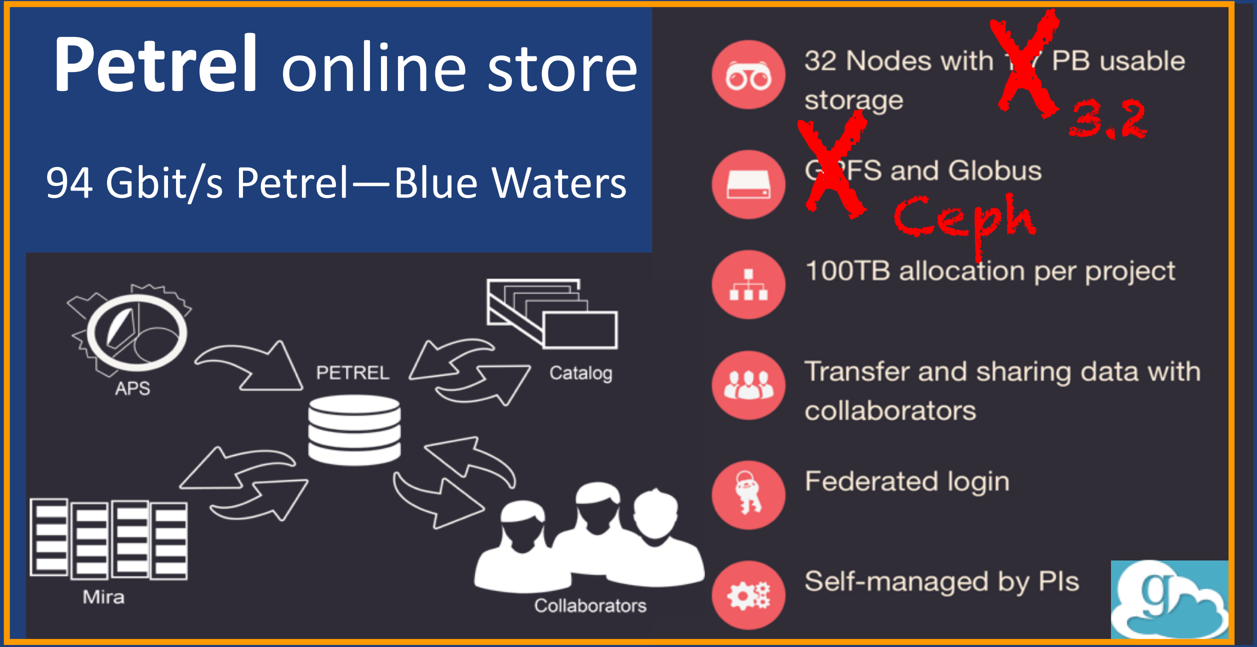
Globus Transfer API Set

- **Doc**
  - https://docs.globus.org/api/transfer/
- **Sample data portal**
  - https://docs.globus.org/modern-research-data-portal/
- **Jupyter notebook**
  - https://github.com/globus/globus-jupyter-notebooks

Globus Auth API Set

- **Doc**
  - – https://docs.globus.org/api/auth/
- **Sample data portal**
  - – https://docs.globus.org/modern-research-data-portal/
- **Native app examples**
  - – https://github.com/globus/native-app-examples

Petrel: A Programmatically Accessible Research Data Service

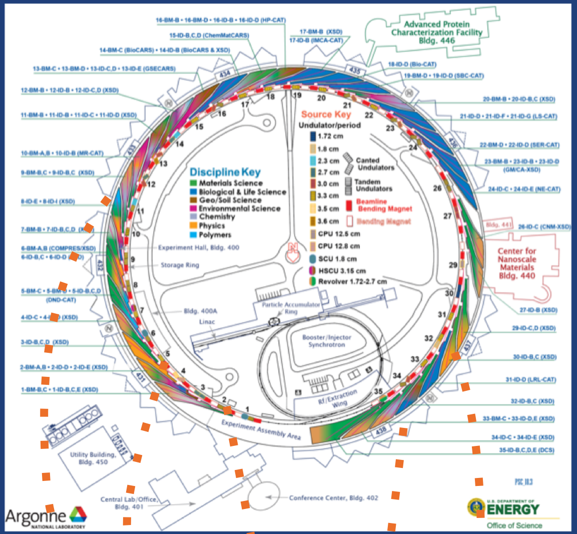


Petrel online store

94 Gbit/s Petrel—Blue Waters

- 32 Nodes with ~~1.7~~ PB usable storage 3.2
- ~~GPFS~~ Ceph and Globus
- 100TB allocation per project
- Transfer and sharing data with collaborators
- Federated login
- Self-managed by PIs

APS
PETREL
Catalog
Mira
Collaborators

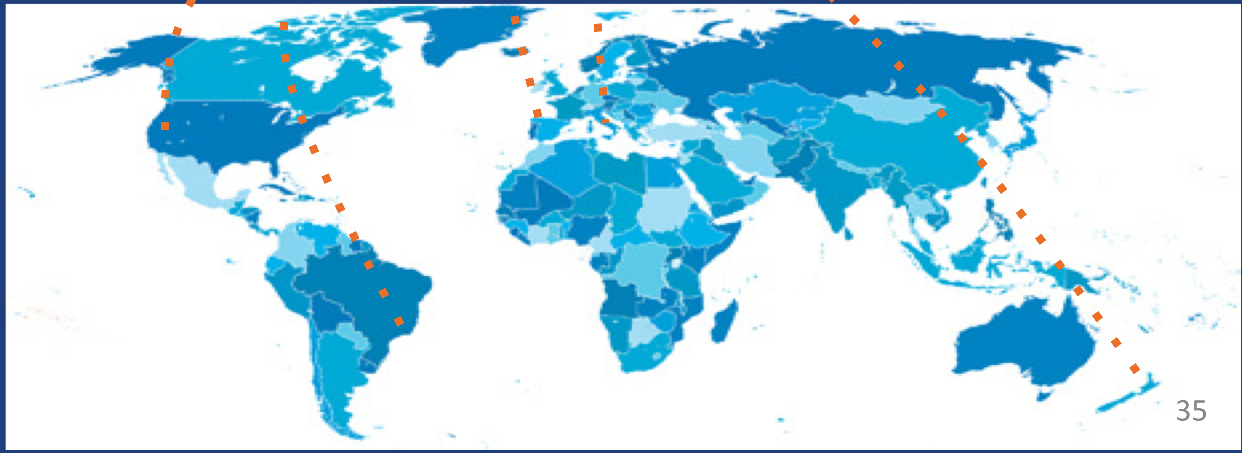


Discipline Key
Source Key
Advanced Protein Characterization Facility Bldg. 446
Center for Nanoscale Materials Bldg. 440
Argonne
ENERGY

PETREL

Data Management and Sharing Pilot

3.2 petabytes
100 Gbps



https://petrel.alcf.anl.gov

A bit of Globus history

U.S. DEPARTMENT OF ENERGY

NSF

THE UNIVERSITY OF CHICAGO

NIH

ALFRED P. SLOAN FOUNDATION 1934

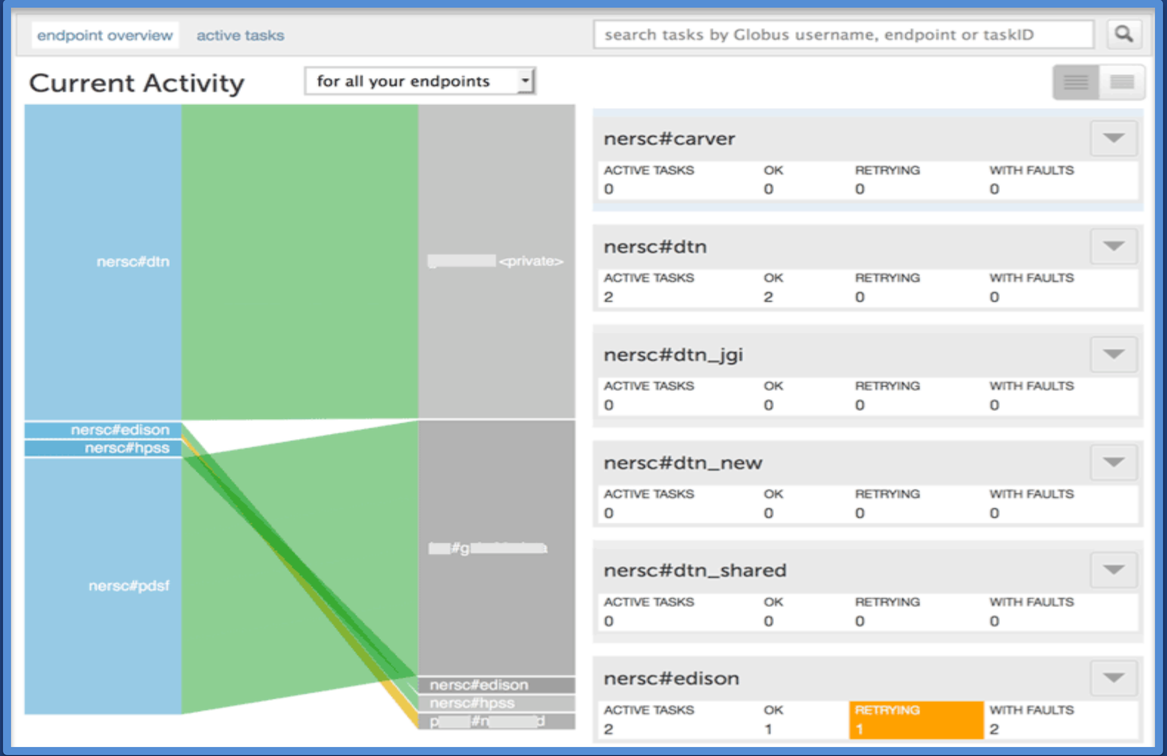
NIST
National Institute of Standards and Technology
U.S. Department of Commerce

Argonne NATIONAL LABORATORY

powered by amazon web services

Globus sustainability model

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



nersc#carver

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

nersc#dtn

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
2	2	0	0

nersc#dtn_jgi

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

nersc#dtn_new

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

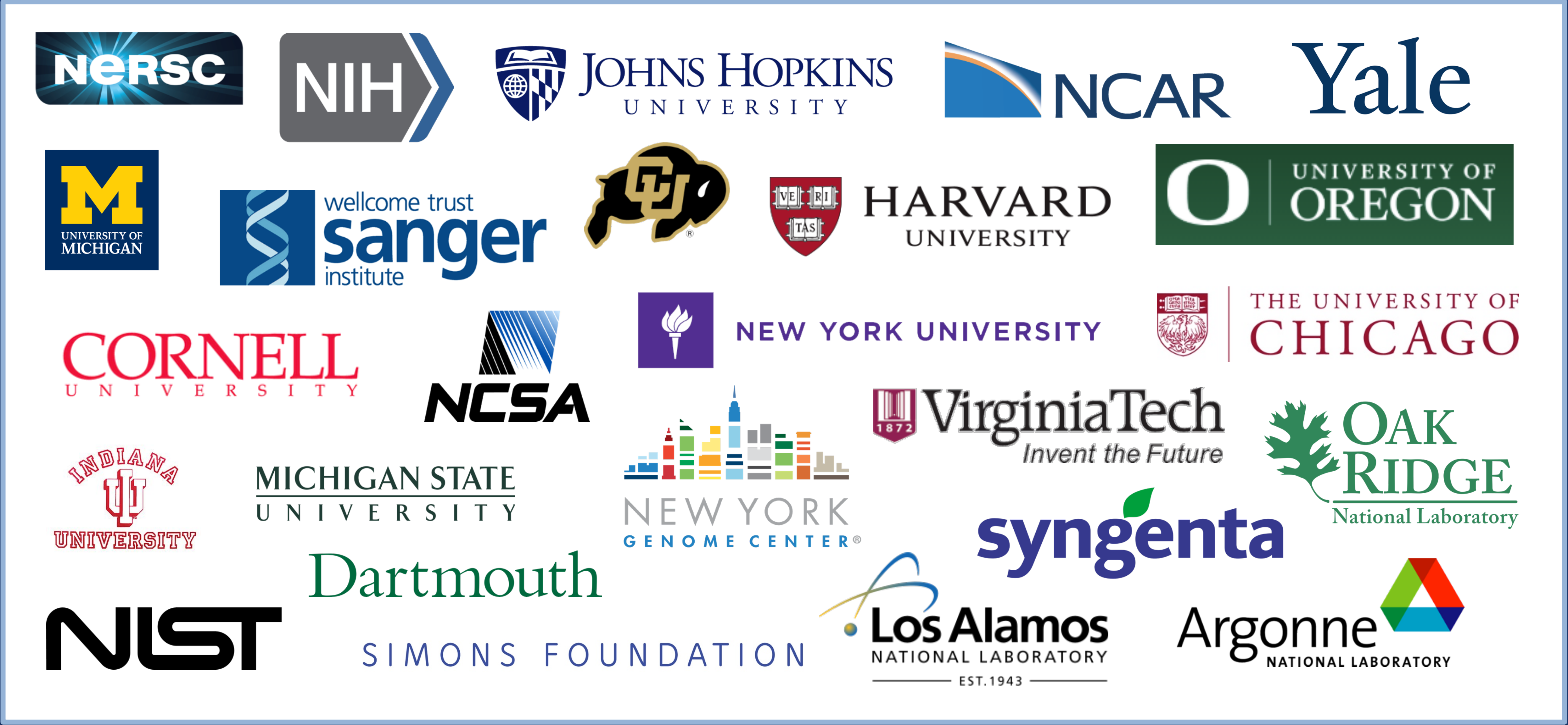
nersc#dtn_shared

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

nersc#edison

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
2	1	1	2

The path to sustainability



Globus by the numbers

1,042
most shared endpoints at a single institution

635+ PB
transferred

73 billion
files processed

1,700
active GCS endpoints

100+
subscribers

100,000+
users

3 months
longest running transfer

18,000
active GCP endpoints

500+
identity providers

1 PB
largest single transfer to date

8,000
active shared endpoints

99.9%
availability

ALCF Globus Resources

- **Documentation**
  - https://www.alcf.anl.gov/user-guides/data-transfer
  - https://www.alcf.anl.gov/user-guides/using-globus
- **Endpoints**
  - Theta: alcf#dtn_theta
  - Mira: alcf#dtn_mira
  - Cetus: alcf#dtn_mira
  - Cooley: alcf#dtn_mira
  - Vesta: alcf#dtn_vesta
  - HPSS: alcf#dtn_hpss

Globus support resources

- **Globus documentation: docs.globus.org**
- **Helpdesk and issue escalation: support@globus.org**
- **Mailing Lists**
  - https://www.globus.org/mailing-lists
- **Customer engagement team**
- **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