Research Data Management using Software-as-a-Service

Vas Vasiliadis
vas@uchicago.edu

THE UNIVERSITY OF CHICAGO

globus

EDUCAUSE 2015

Presentation material available at

globus.org/events/educause-2015

bit.ly/educause2015

Thank you to our sponsors!

U.S. DEPARTMENT OF ENERGY

NSF

NATIONAL INSTITUTES OF HEALTH

ALFRED P. SLOAN FOUNDATION 1934

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

powered by amazon web services

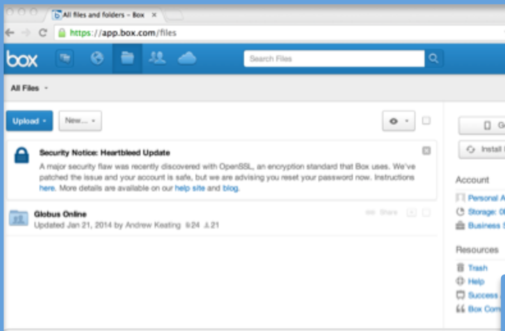
Agenda

- Research data management challenges
- Globus: a high-level flyover
- Accelerating and streamlining collaboration: transfer and sharing
- Enhancing reproducibility and discoverability: data publication
- Our sustainability challenge
- Campus deployment, security overview
- Leveraging the Globus platform

Who are you?

Research data management scenarios and challenges

"I need to easily, quickly, & reliably move portions of my data to other locations."



All files and folders – Box

https://app.box.com/files

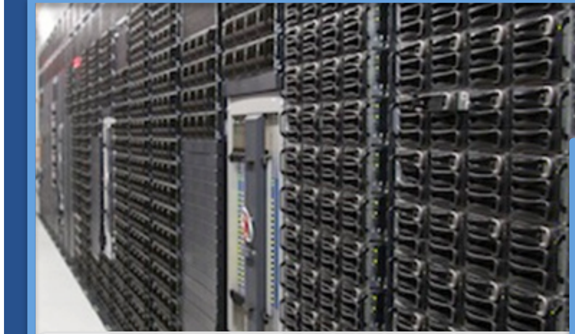
Security Notice: Heartbleed Update

A major security flaw was recently discovered with OpenSSL, an encryption standard that Box uses. We've patched the issue and your account is safe, but we are advising you reset your password now. Instructions here. More details are available on our help site and blog.

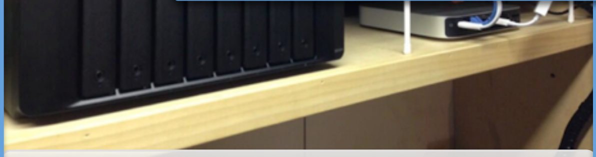
Globus Online

Updated Jan 21, 2014 by Andrew Keating 24 21

Campus Home Filesyster



Research Computing HPC Cluste



Lab Server

Personal Laptop

Desktop Workstation

XSEDE Resource

amazon web services S3

red cloud

Windows Azure

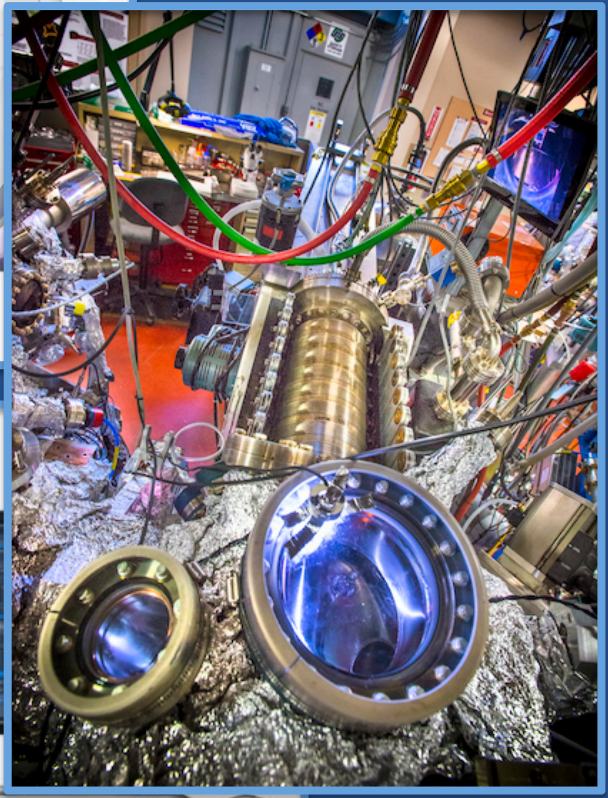
Google Cloud Platform

Public Cloud

“I need to get data from a scientific instrument to my analysis system.”

MRI

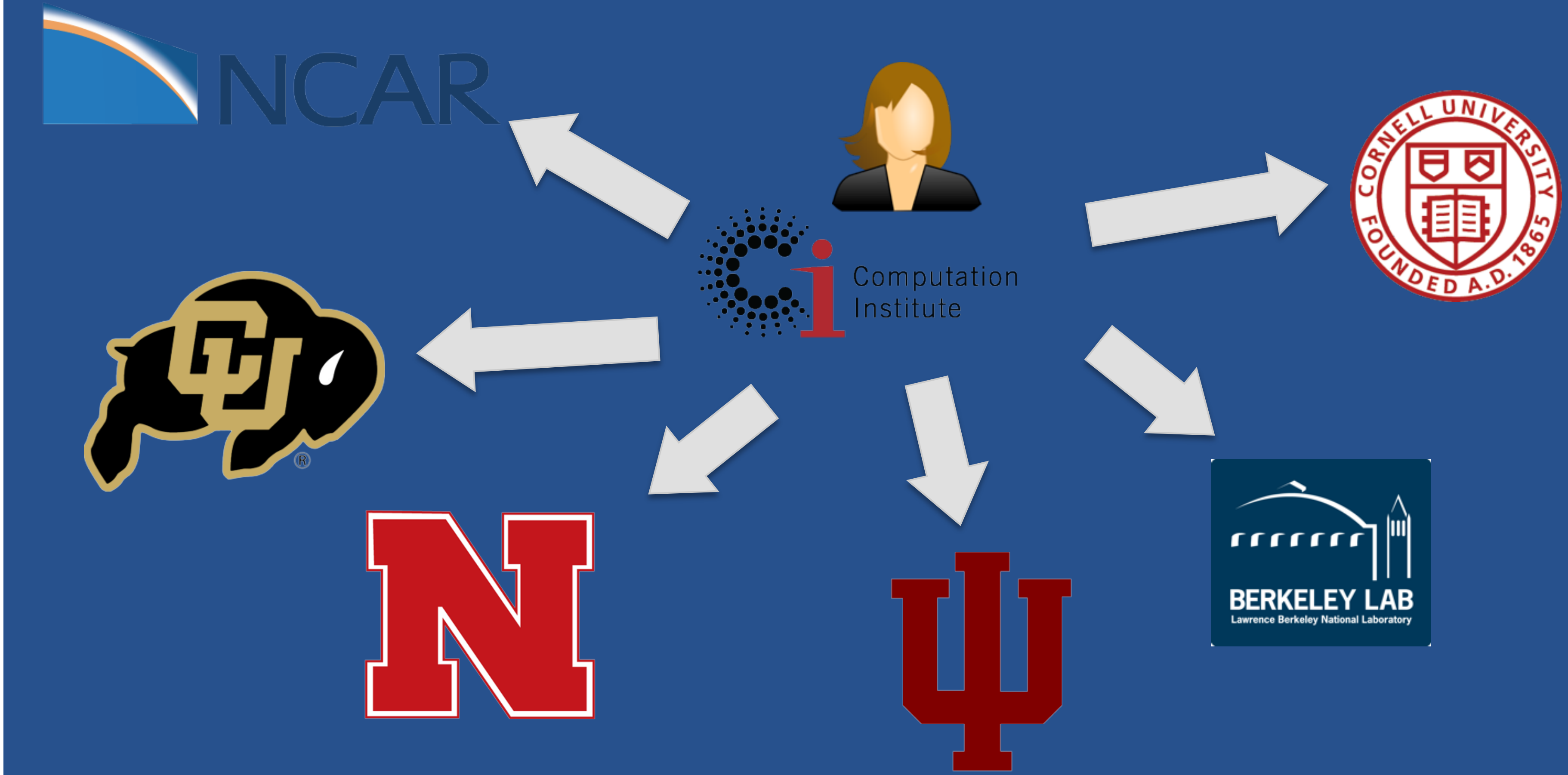
Advanced Light Source



Next Gen Sequencer

Light Sheet Microscope

"I need to easily and securely share my data with my colleagues at other institutions."


NCAR
Computation Institute
CORNELL UNIVERSITY FOUNDED A.D. 1865
CU
N
BERKELEY LAB
Lawrence Berkeley National Laboratory

“I need to publish my data so others can find/use/validate/reproduce it.”

Scholarly Publication

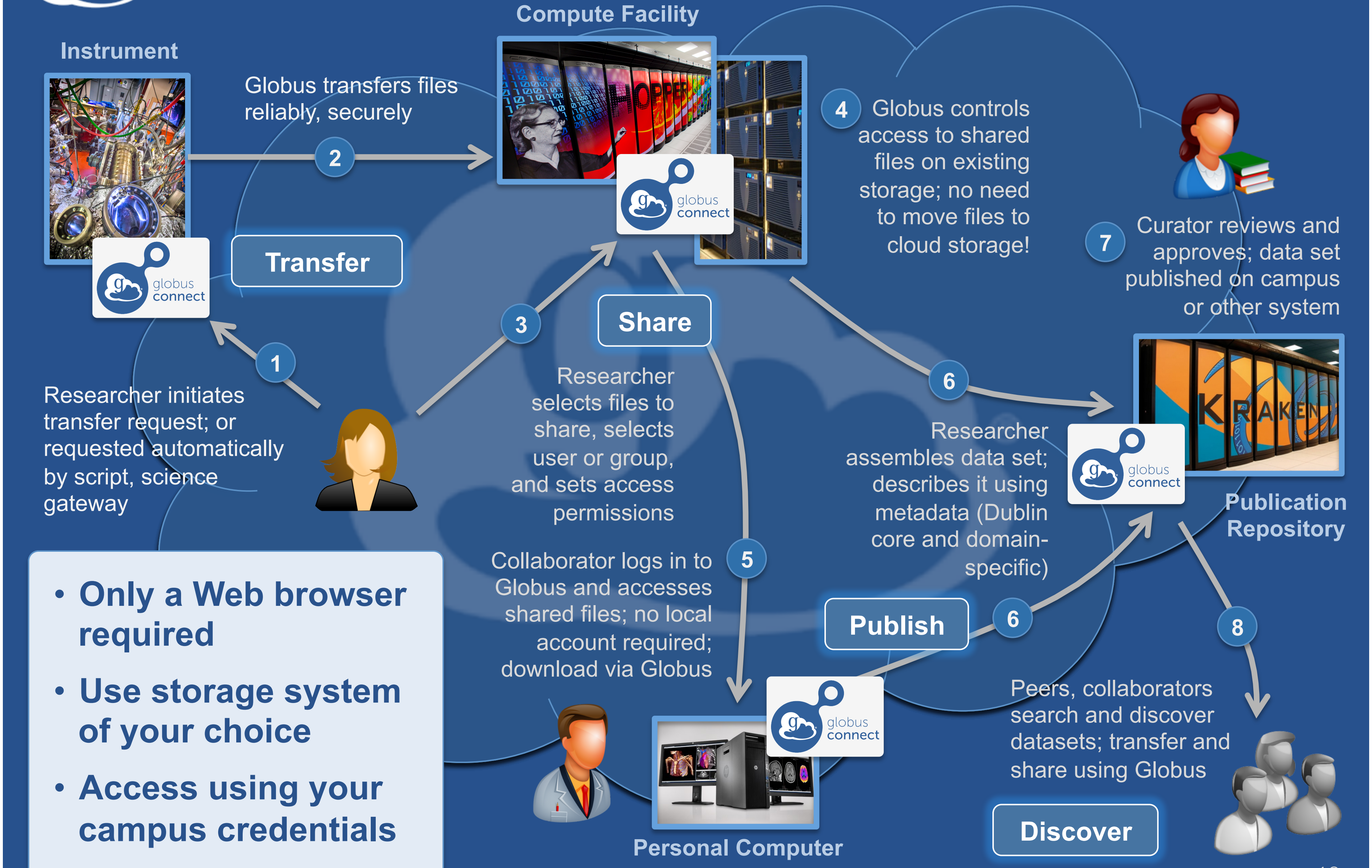
Reference Dataset

Research Community Collaboration

Research data management today

Index?

Globus and the research data lifecycle



Instrument

Compute Facility

Personal Computer

Publication Repository

Transfer

Share

Publish

Discover

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

- **Only a Web browser required**
- **Use storage system of your choice**
- **Access using your campus credentials**

Globus delivers…

Big data transfer, sharing, publication, and discovery…

…directly from your own storage systems…

...via software-as-a-service

Globus is SaaS

- **Easy to access via Web browser**
  - Command line, REST interfaces for flexible automation and integration
- **New features automatically available**
- **Reduced IT operational costs**
  - Small local footprint (Globus Connect)
  - Consolidated support and troubleshooting

Our focus: User Experience

flickr ...for your photos

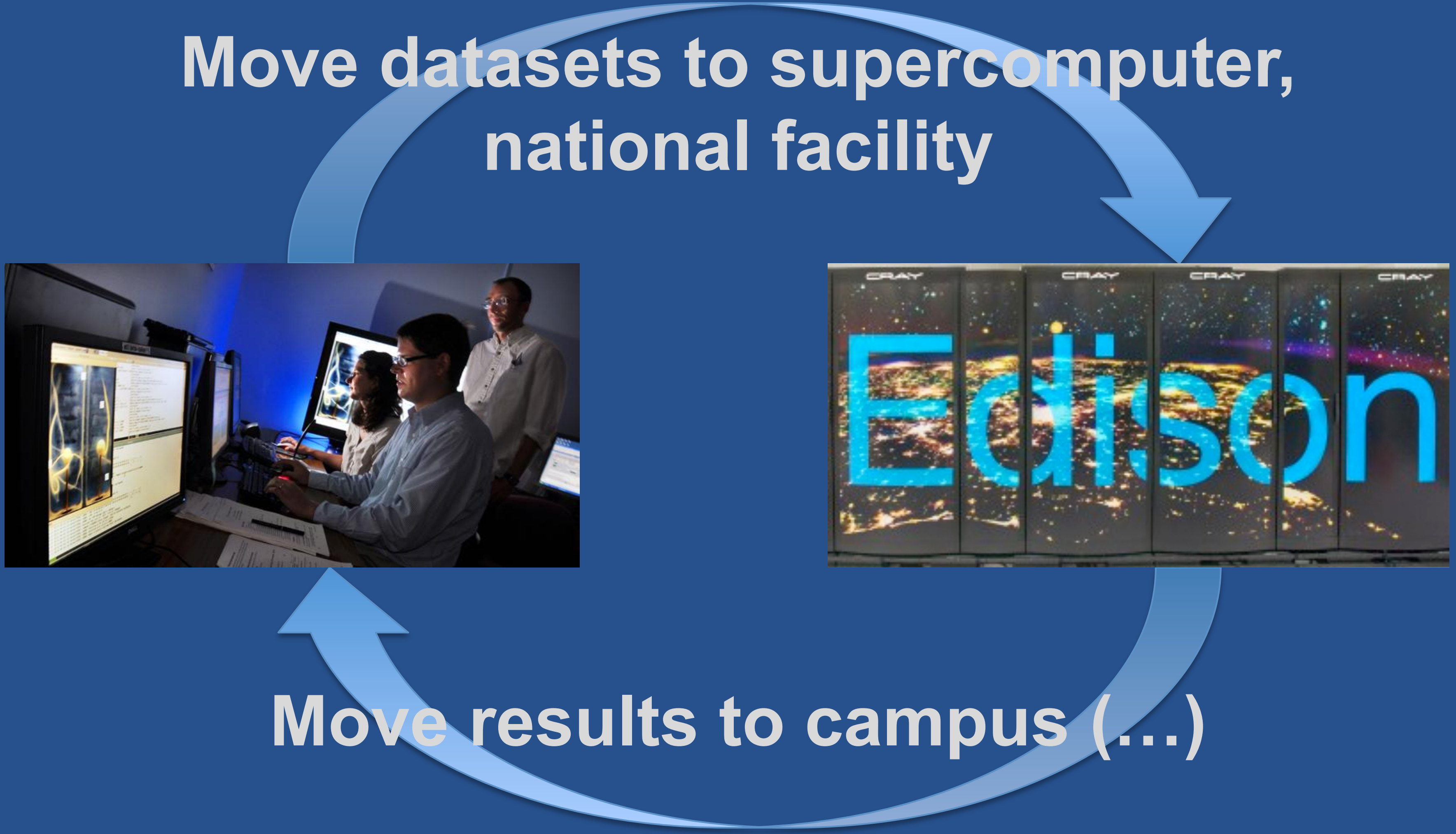
Google ...for your office docs

NETFLIX ...for your entertainment

globus ...for your research data

Accessing Globus and Moving Data

Example: Scaling up



Move datasets to supercomputer, national facility

Move results to campus (…)

Sign up & transfer files

1. Go to: www.globus.org/signup
2. Create your Globus account
3. Validate e-mail address
4. Optional: Login with your campus/ InCommon identity
5. Install Globus Connect Personal
6. Move files from vas#ebs endpoint to your laptop

Sharing Data

Lowering collaboration overhead

- **Grant collaborators access to data on systems without requiring local accounts**
- **No need to replicate or move data to separate system/cloud just for sharing**
- **Researchers manage “virtual” ACLs…**
- **Respect local system access controls**

Share files

1. **Join the “Tutorial Users” groups**
   - Go to “Groups”, search for “tutorial”
   - Select group from list, click “Join Group”
2. **Create a shared endpoint on your laptop**
3. **Grant your neighbor permissions on your shared endpoint**
4. **Access your neighbor’s shared endpoint**

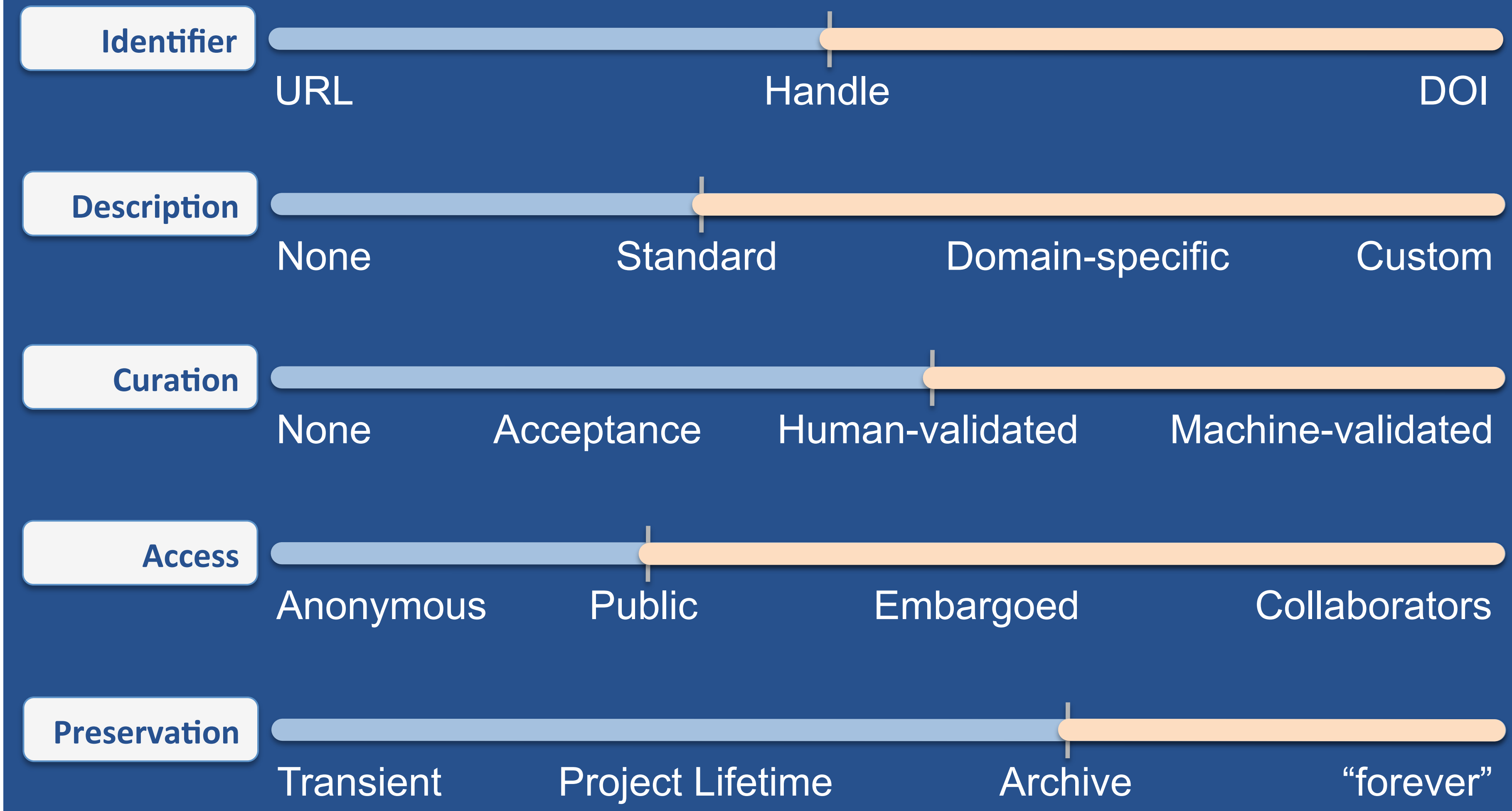
Group Management

Exercise 3: Create/configure group

1. **Create a group**
   - Go to globus.org/groups
   - Click “Create New Group”
   - Enter the group name and a short description
   - Set visibility to “all Globus members”
2. **Configure your group policies**
   - Select your group and click the “Settings” tab
   - Set requests to “a logged in Globus user”
   - Set approvals to “automatically if all policies are met”
3. **Ask your neighbor to join your group**
4. **Grant permissions to the group on your shared endpoint**
5. **Confirm your neighbor can access your shared endpoint**

Enhancing reproducibility and discoverability

Globus data publication framework



Identifier

URL Handle DOI

Description

None Standard Domain-specific Custom

Curation

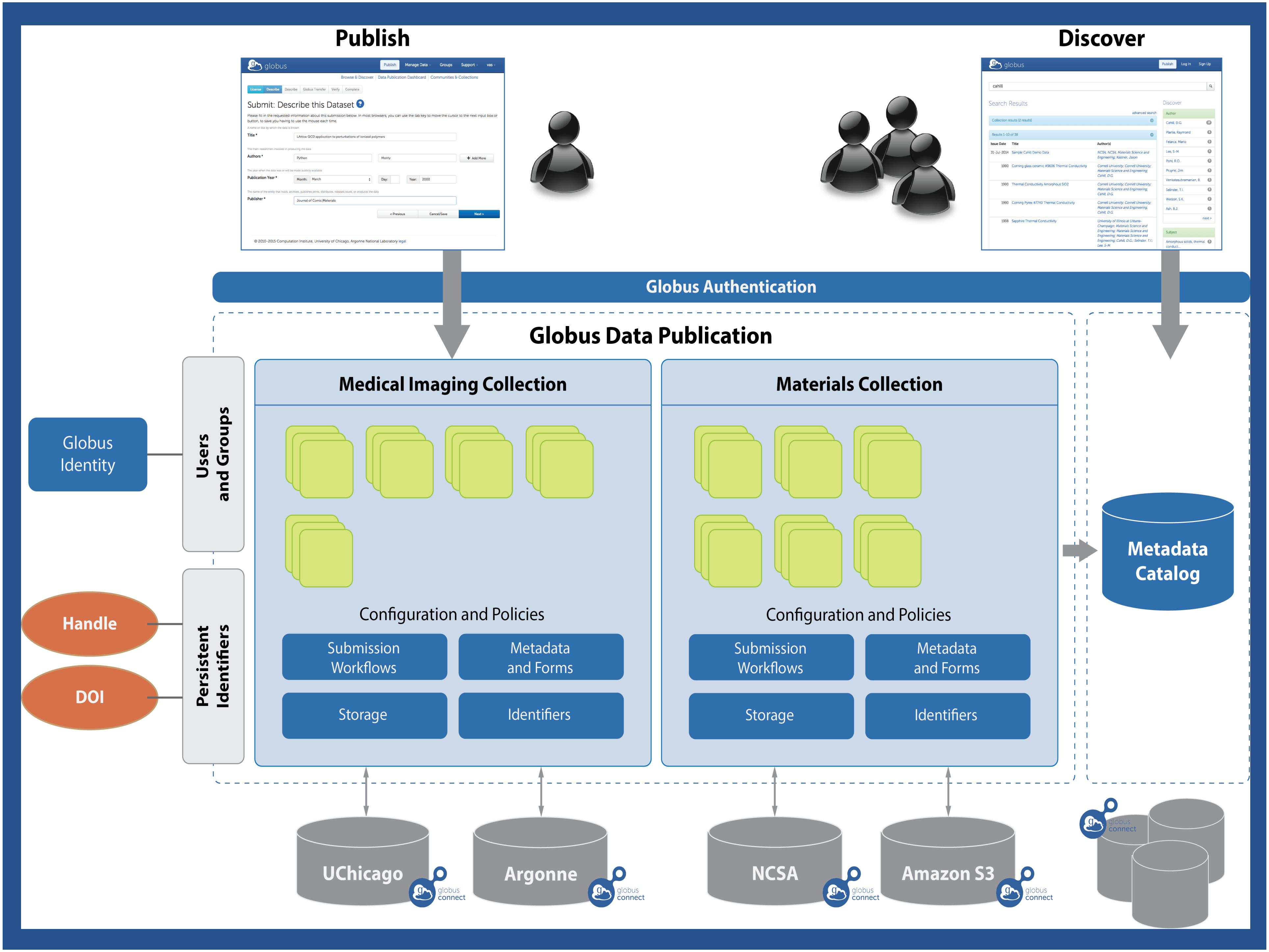
None Acceptance Human-validated Machine-validated

Access

Anonymous Public Embargoed Collaborators

Preservation

Transient Project Lifetime Archive “forever”



globus — Publish | Manage Data | Groups | Support | vas

Browse & Discover | Data Publication Dashboard | Communities & Collections

License | Describe | Describe | Globus Transfer | Verify | Complete

Submit: Describe this Dataset

Please fill in the requested information about this submission below. In most browsers, you can use the tab key to move the cursor to the next input box or button, to save you having to use the mouse each time.

A name or title by which the data is known

Title * LAttice QCD application to perturbations of ionized polymers

The main researchers involved in producing the data

Authors * Python | Monty | + Add More

The year when the data was or will be made publicly available

Publication Year * Month: March | Day: | Year: 2000

The name of the entity that holds, archives, publishes prints, distributes, releases,issues, or produces the data

Publisher * Journal of Comic Materials

< Previous | Cancel/Save | Next >

© 2010-2015 Computation Institute, University of Chicago, Argonne National Laboratory legal

globus — Publish | Log In | Sign Up

cahill

Search Results

advanced search

Collection results (2 results)

Results 1-10 of 38

Issue Date	Title	Author(s)
31-Jul-2014	Sample Cahill Demo Data	NCSA; NCSA; Materials Science and Engineering; Kastner, Jason
1990	Corning glass ceramic #9606 Thermal Conductivity	Cornell University; Cornell University; Materials Science and Engineering; Cahill, D.G.
1990	Thermal Conductivity Amorphous SiO2	Cornell University; Cornell University; Materials Science and Engineering; Cahill, D.G.
1990	Corning Pyrex #7740 Thermal Conductivity	Cornell University; Cornell University; Materials Science and Engineering; Cahill, D.G.
1998	Sapphire Thermal Conductivity	University of Illinois at Urbana-Champaign; Materials Science and Engineering; Materials Science and Engineering; Materials Science and Engineering; Cahill, D.G.; Selinder, T.I.; Lee, S-M

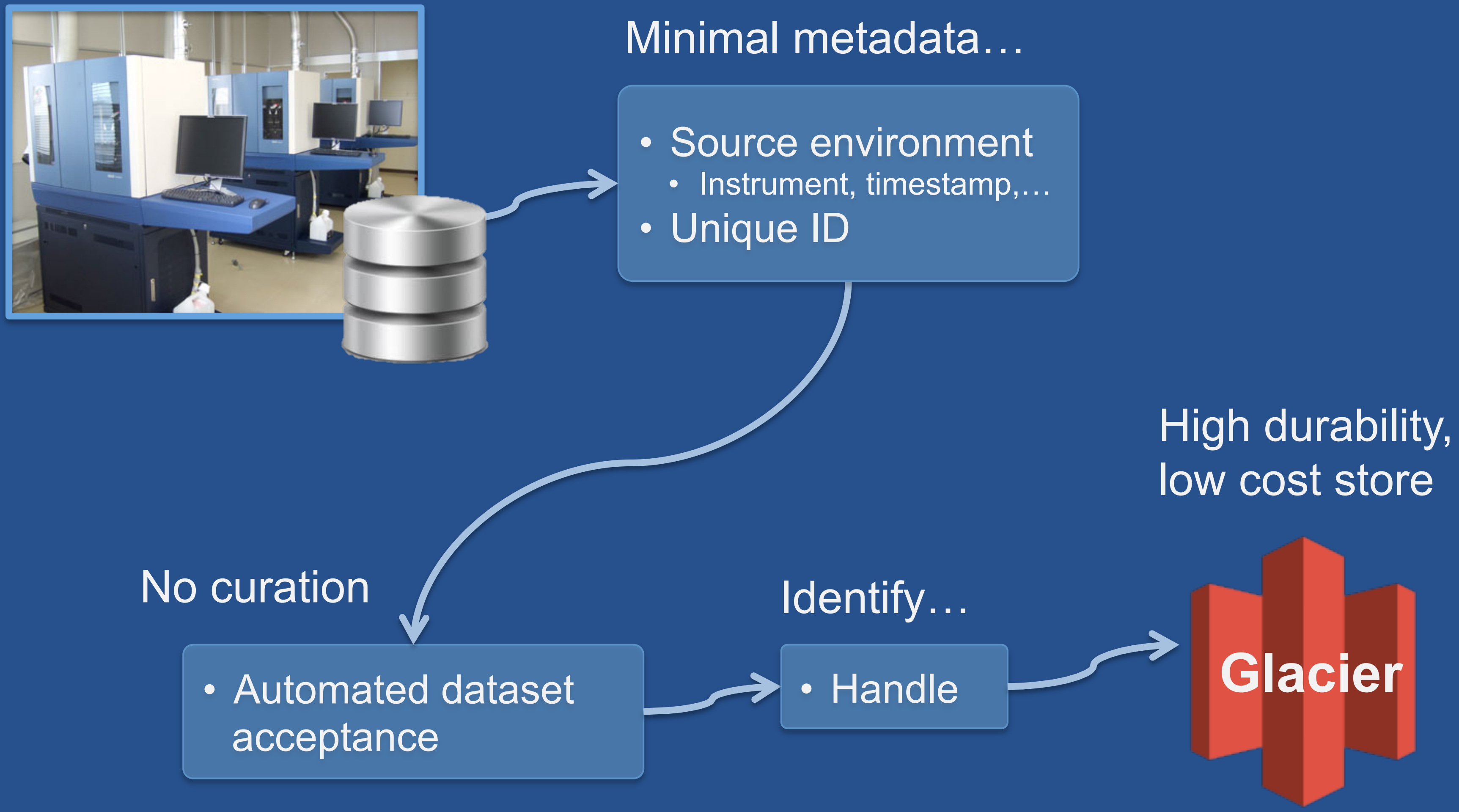
Discover

Author	
Cahill, D.G.	12
Plante, Raymond	6
Felarca, Mario	5
Lee, S-M	5
Pohl, R.O.	3
Pruyne, Jim	3
Venkatasubramanian, R.	3
Selinder, T.I.	2
Watson, S.K.	2
Ash, B.J.	1

next >

Subject	
Amorphous solids, thermal conduct...	3

Raw NGS output



Minimal metadata…

- Source environment
  - Instrument, timestamp,…
- Unique ID

No curation

- Automated dataset acceptance

Identify…

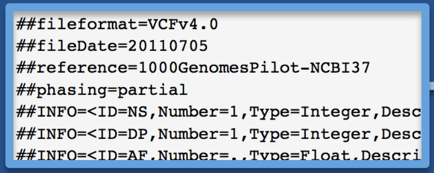
- Handle

High durability, low cost store

Upstream analysis

Campus HPC

globus **genomics**



```
##fileformat=VCFv4.0
##fileDate=20110705
##reference=1000GenomesPilot-NCBI37
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Desc
##INFO=<ID=DP,Number=1,Type=Integer,Desc
##INFO=<ID=AF,Number=.,Type=Float,Descri
```

Processing metadata…

- Pipeline description
- Tool parameters
- Exec environment

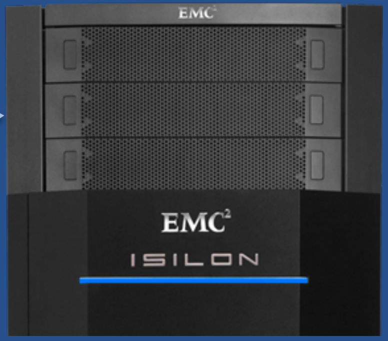
Automated curation

- Machine validated
- Exception review

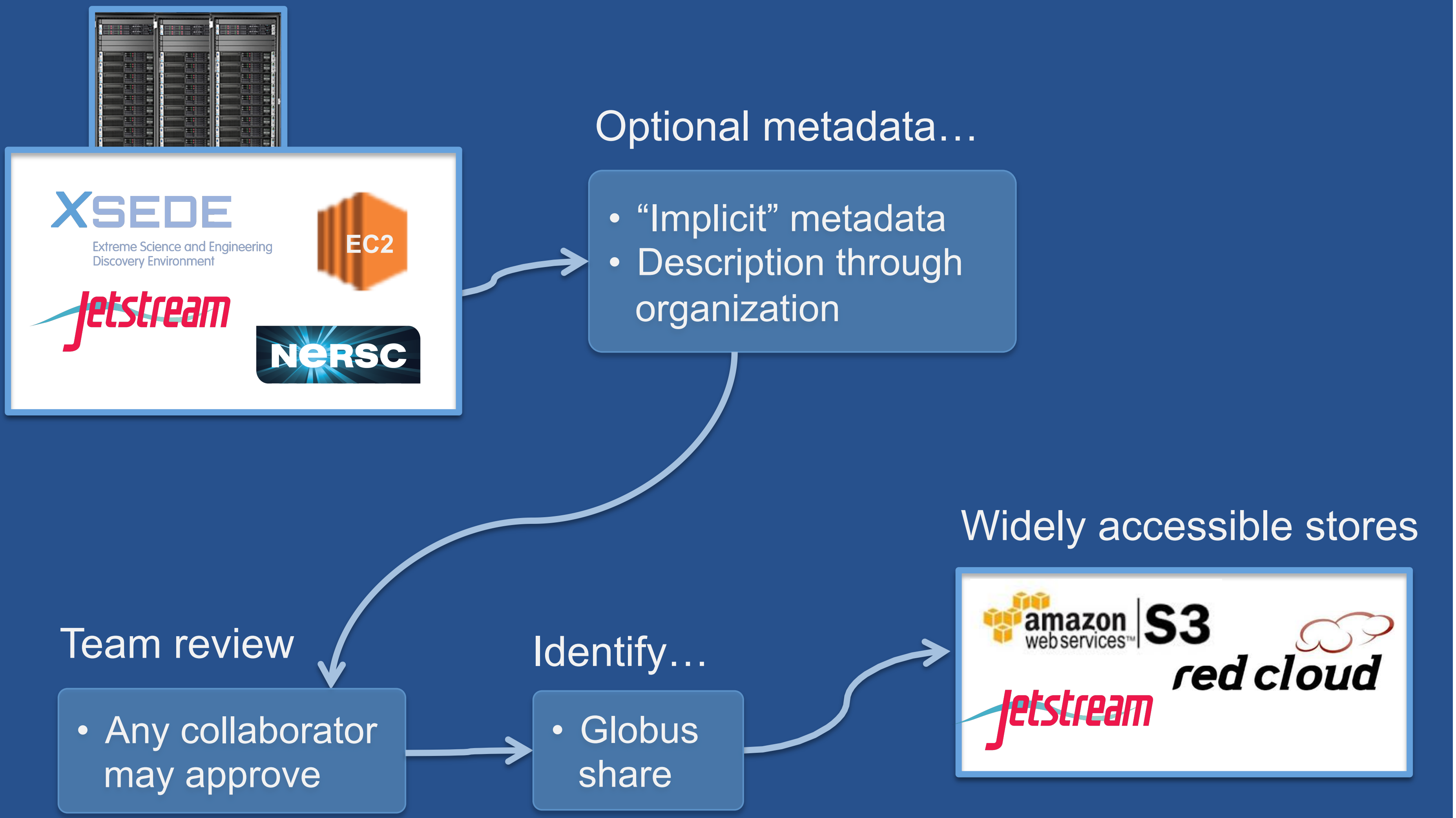
Identify…

- URL

Moderate durability/cost



Downstream analysis



Optional metadata…

- “Implicit” metadata
- Description through organization

Team review

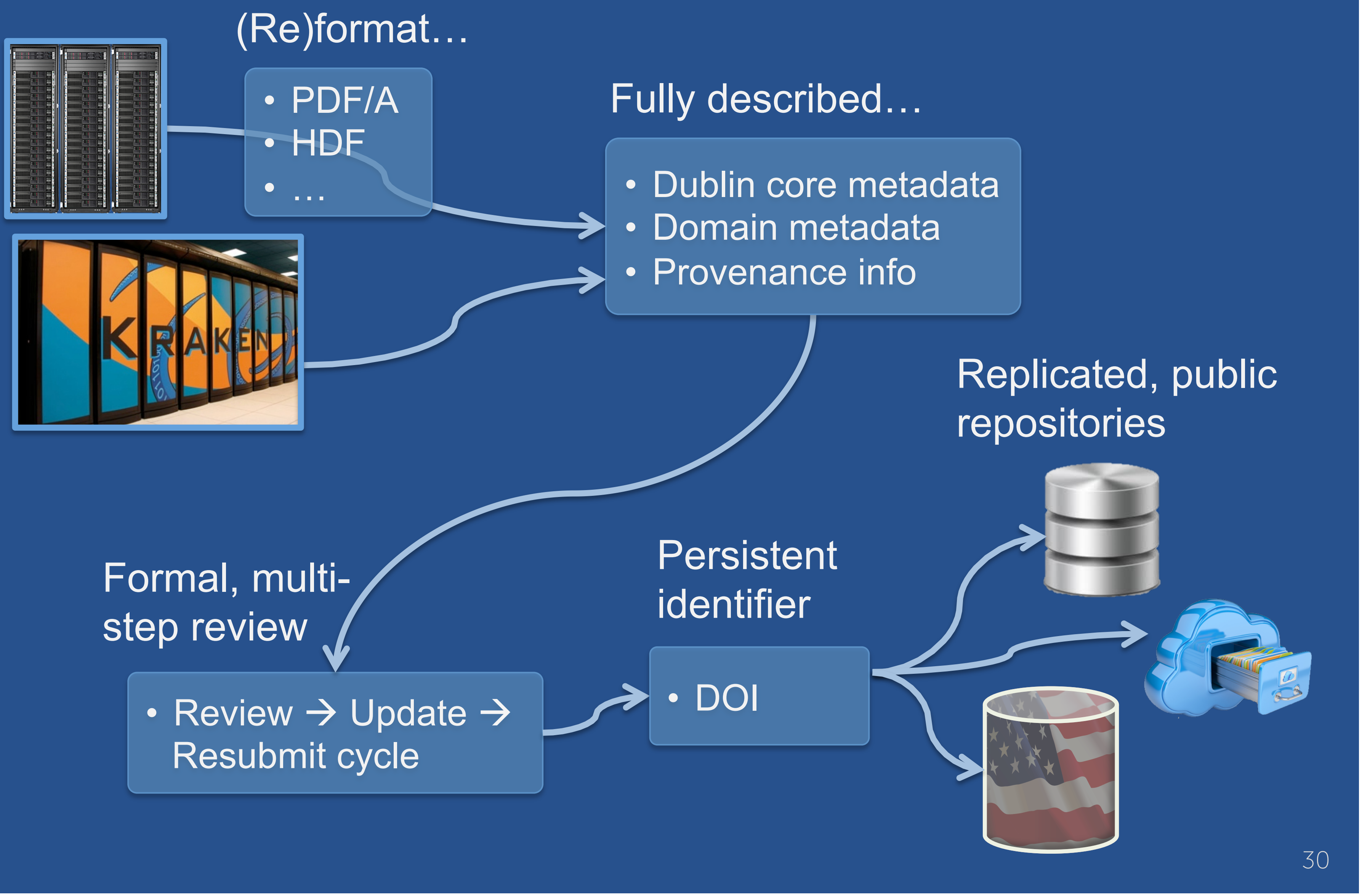
- Any collaborator may approve

Identify…

- Globus share

Widely accessible stores

Peer reviewed paper



(Re)format…

- PDF/A
- HDF
- …

Fully described…

- Dublin core metadata
- Domain metadata
- Provenance info

Replicated, public repositories

Formal, multi-step review

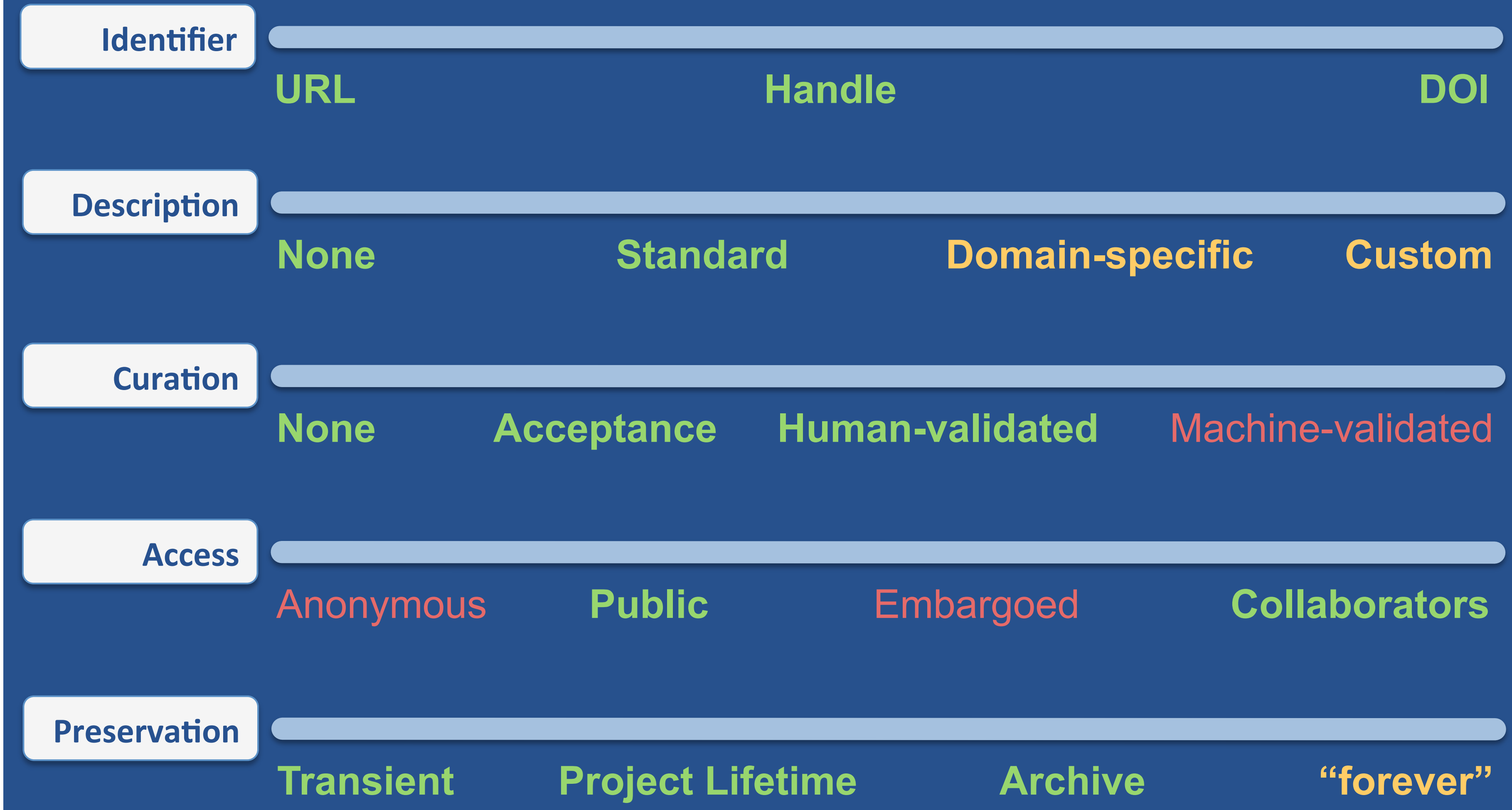
- Review → Update → Resubmit cycle

Persistent identifier

- DOI

Globus publication - Initial release

🟩 Supported in GA release 🟨 Consulting support 🟥 Planned



Identifier

URL Handle DOI

Description

None Standard Domain-specific Custom

Curation

None Acceptance Human-validated Machine-validated

Access

Anonymous Public Embargoed Collaborators

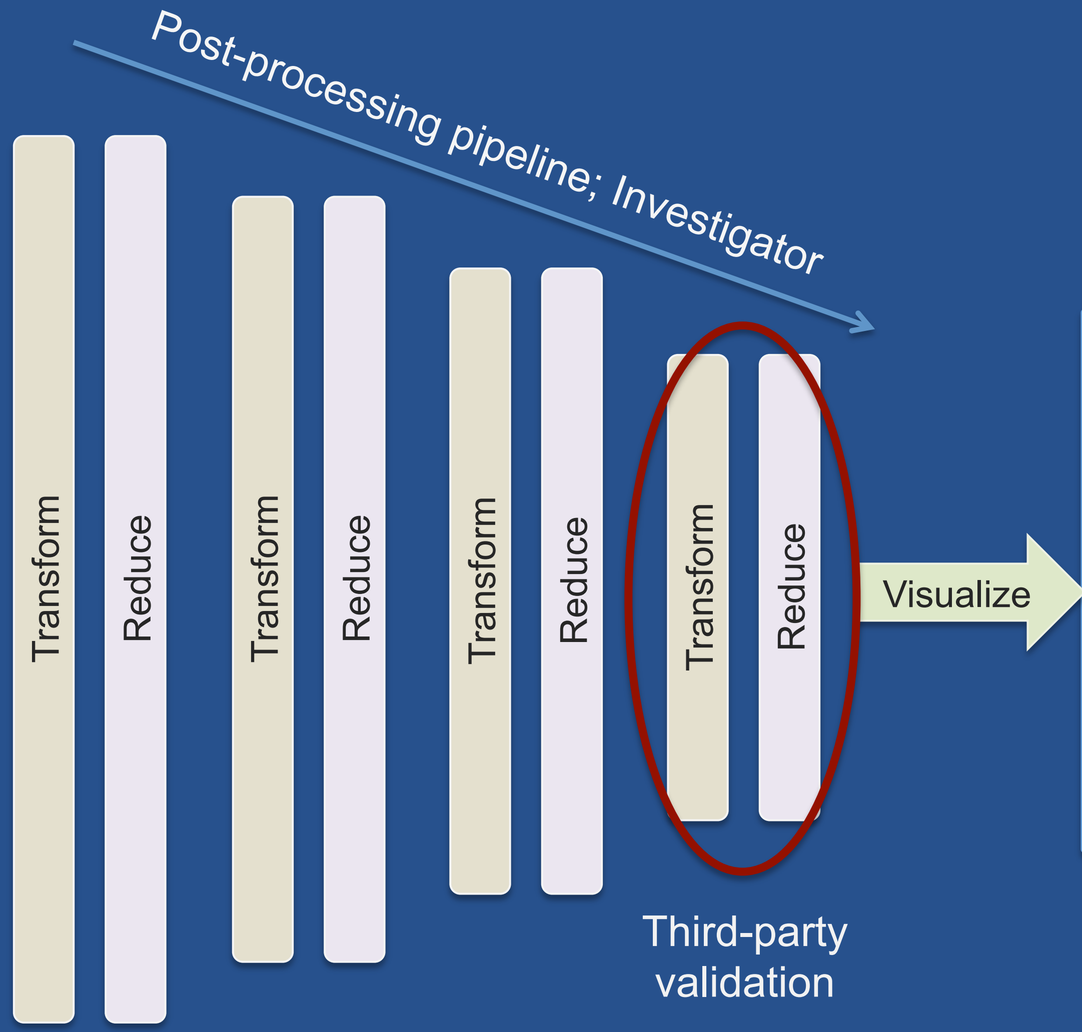
Preservation

Transient Project Lifetime Archive “forever”

Publish a dataset

1. Go to trial.publish.globus.org
2. Log in, click “Submit a New Dataset”
3. Select either of the Open Trial collections and continue
4. Accept the license terms
5. Enter required metadata to describe the dataset
6. Assemble data set from the vas#ebs endpoint (or your own laptop if you installed Globus Connect Personal)
7. Complete the workflow and submit
8. Curators (a.k.a. presenters) will “review” your submission and publish
9. Search for your published dataset and browse the data

Reproducibility example



Post-processing pipeline; Investigator

Transform

Reduce

Transform

Reduce

Transform

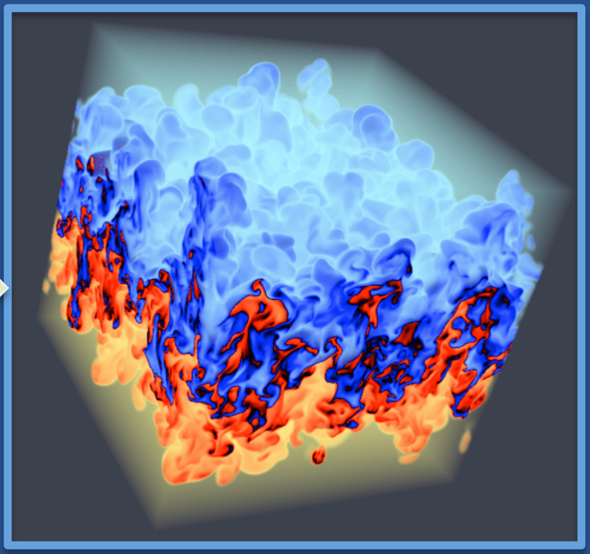
Reduce

Transform

Reduce

Visualize

Third-party validation



Source: M. Hutchinson, R. Rosner, University of Chicago; Argonne; Image: UNM

Reproducibility example

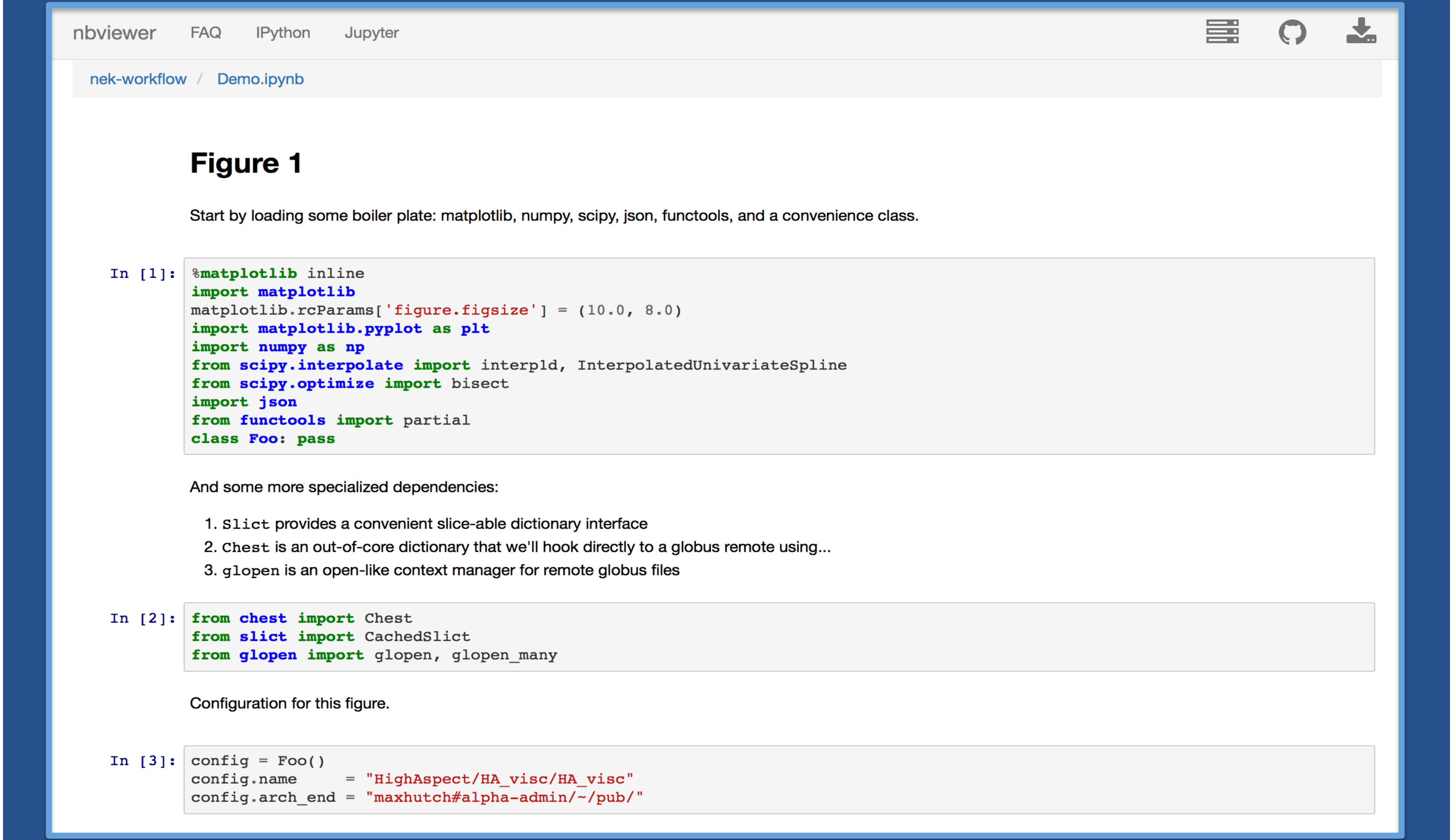


Figure 1

Start by loading some boiler plate: matplotlib, numpy, scipy, json, functools, and a convenience class.

```
In [1]: %matplotlib inline
        import matplotlib
        matplotlib.rcParams['figure.figsize'] = (10.0, 8.0)
        import matplotlib.pyplot as plt
        import numpy as np
        from scipy.interpolate import interp1d, InterpolatedUnivariateSpline
        from scipy.optimize import bisect
        import json
        from functools import partial
        class Foo: pass
```

And some more specialized dependencies:

1. `Slict` provides a convenient slice-able dictionary interface
2. `Chest` is an out-of-core dictionary that we'll hook directly to a globus remote using...
3. `glopen` is an open-like context manager for remote globus files

```
In [2]: from chest import Chest
        from slict import CachedSlict
        from glopen import glopen, glopen_many
```

Configuration for this figure.

```
In [3]: config = Foo()
        config.name      = "HighAspect/HA_visc/HA_visc"
        config.arch_end = "maxhutch#alpha-admin/~/pub/"
```

Source: M. Hutchinson, R. Rosner, University of Chicago

Repository planning

NEW YORK UNIVERSITY

Mix-Model: Technology meets Pedagogy

- **Central IT provides infrastructure**
  - Storage, Computing, Cluster, Servers...
- **Library responsible for data stewardship**
  - Collection, Acquisition, Search & Discovery, Metadata, Preservation...
- **Staff: technologists + librarians and subject specialists**

Source: H. Mistry, New York University

Globus: today and tomorrow

Globus today…

4
major services

118 PB
transferred

20 billion
files processed

31,000
registered users

13
national labs use Globus

10,000
active endpoints

~350
active daily users

99.95%
uptime

35+
institutional subscribers

1 PB
largest single transfer to date

3 months
longest continuously managed transfer

130
federated campus identities

We are a non-profit, delivering a production-grade service to the non-profit research community

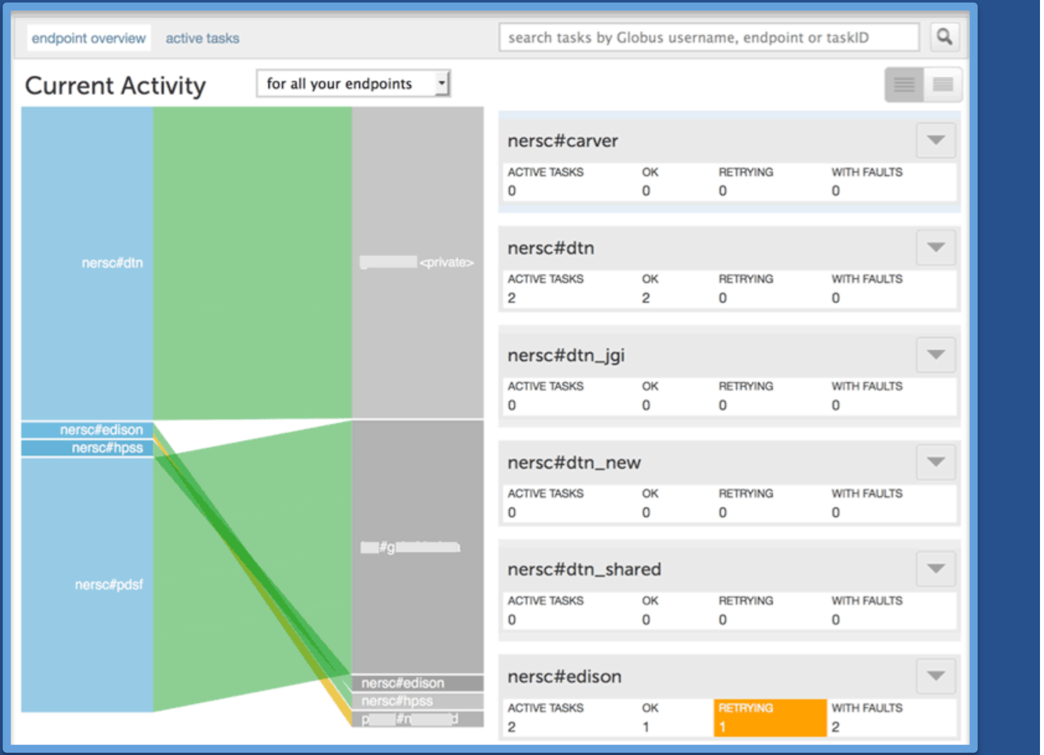
We are a non-profit, delivering a production-grade service to the non-profit research community

Our challenge:

Sustainability

Globus Provider Subscriptions

- **Globus Provider Plan**
  - Shared endpoints
  - Data publication
  - Amazon S3 endpoints
  - Management console
  - Usage reporting
  - Priority support
  - Application integration
- **Branded Web Site**
- **Alternate Identity Provider (InCommon is standard)**
- **Mass Storage System optimization**



Current Activity

for all your endpoints

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

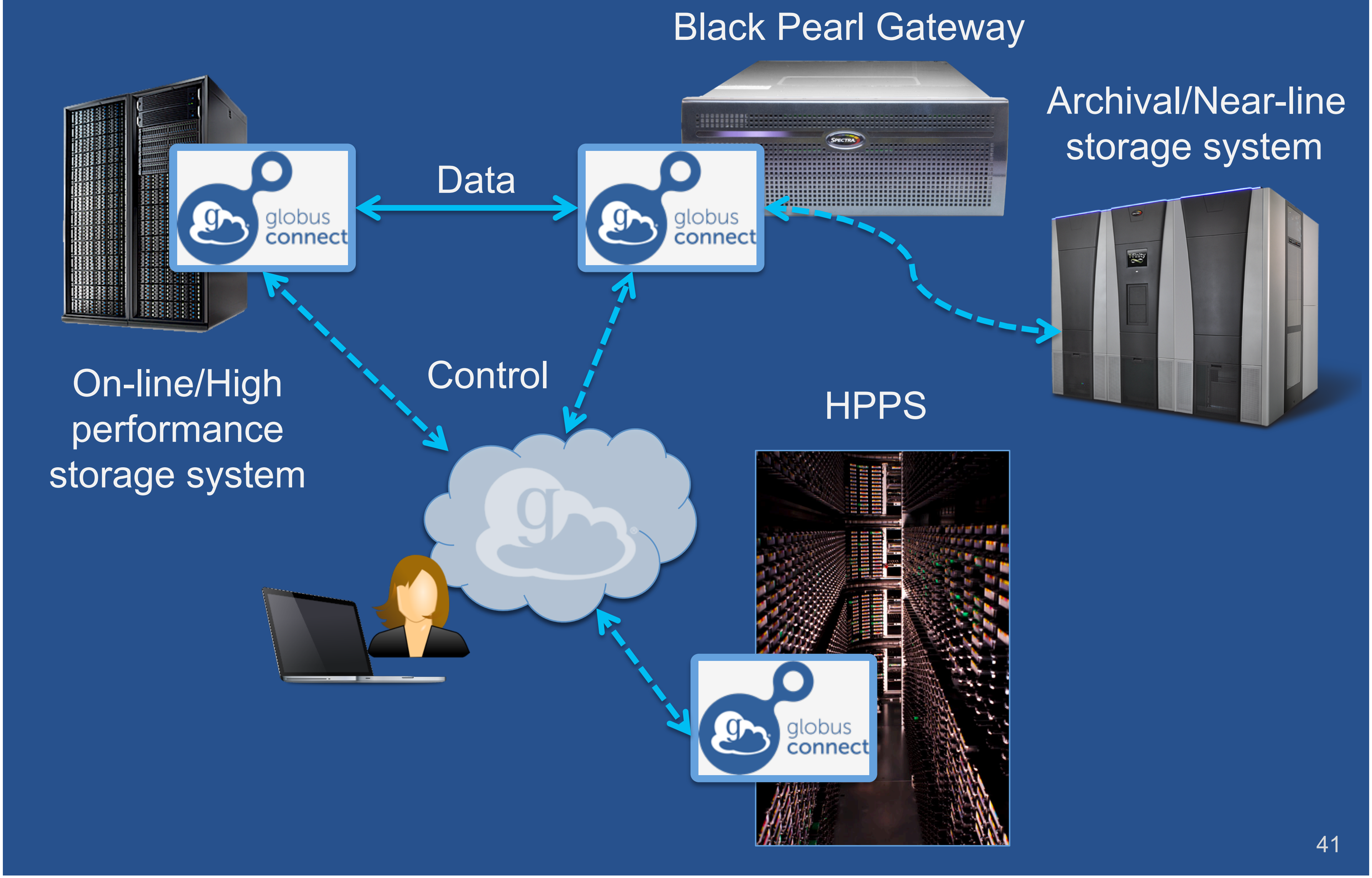
ACTIVE TASKS	OK	RETRYING	WITH FAULTS
0	0	0	0

nersc#edison

ACTIVE TASKS	OK	RETRYING	WITH FAULTS
2	1	1	2

globus.org/provider-plans

Bridging the storage hierarchy

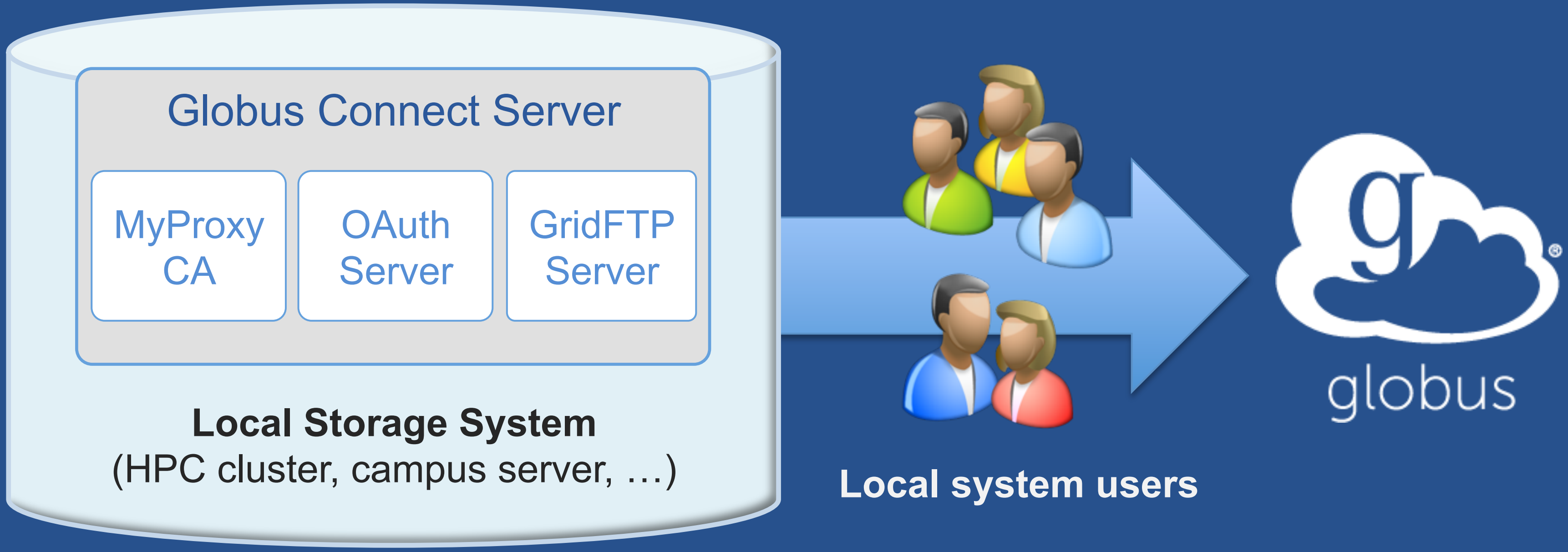


Demonstration:
Globus management console

Demonstration:
Bridging to Cloud Storage
- Amazon S3: supported
- Ceph: coming soon

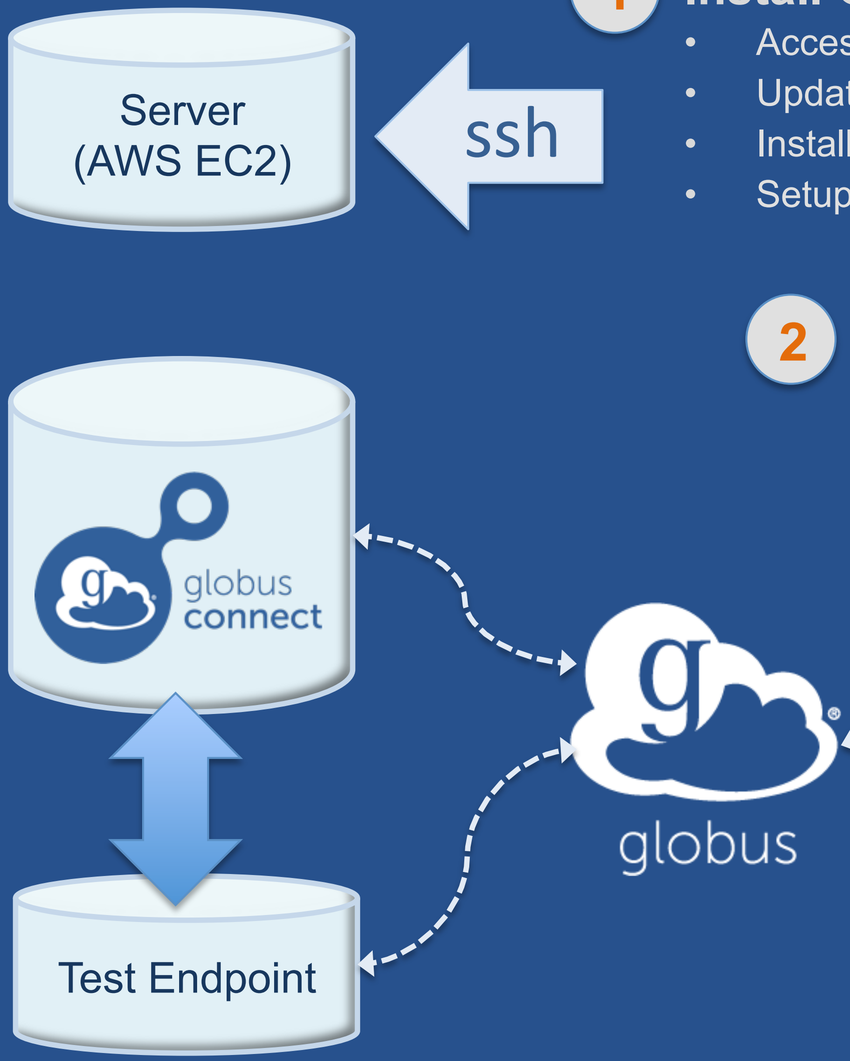
Campus Deployment Overview

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

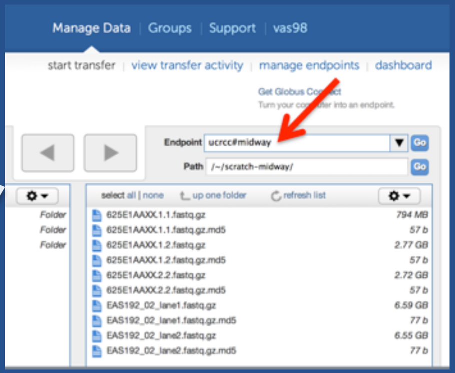
Standard package installation



1 **Install Globus Connect Server**

- Access server with root privileges
- Update package repos
- Install packages
- Setup Globus Connect Server

2 **Log into Globus as end user/ researcher**


Manage Data | Groups | Support | vas98
start transfer | view transfer activity | manage endpoints | dashboard
Get Globus Connect
Turn your computer into an endpoint.
Go
Go
select all | none
up one folder
refresh list

Endpoint ucrcc#midway

Path /~/scratch-midway/

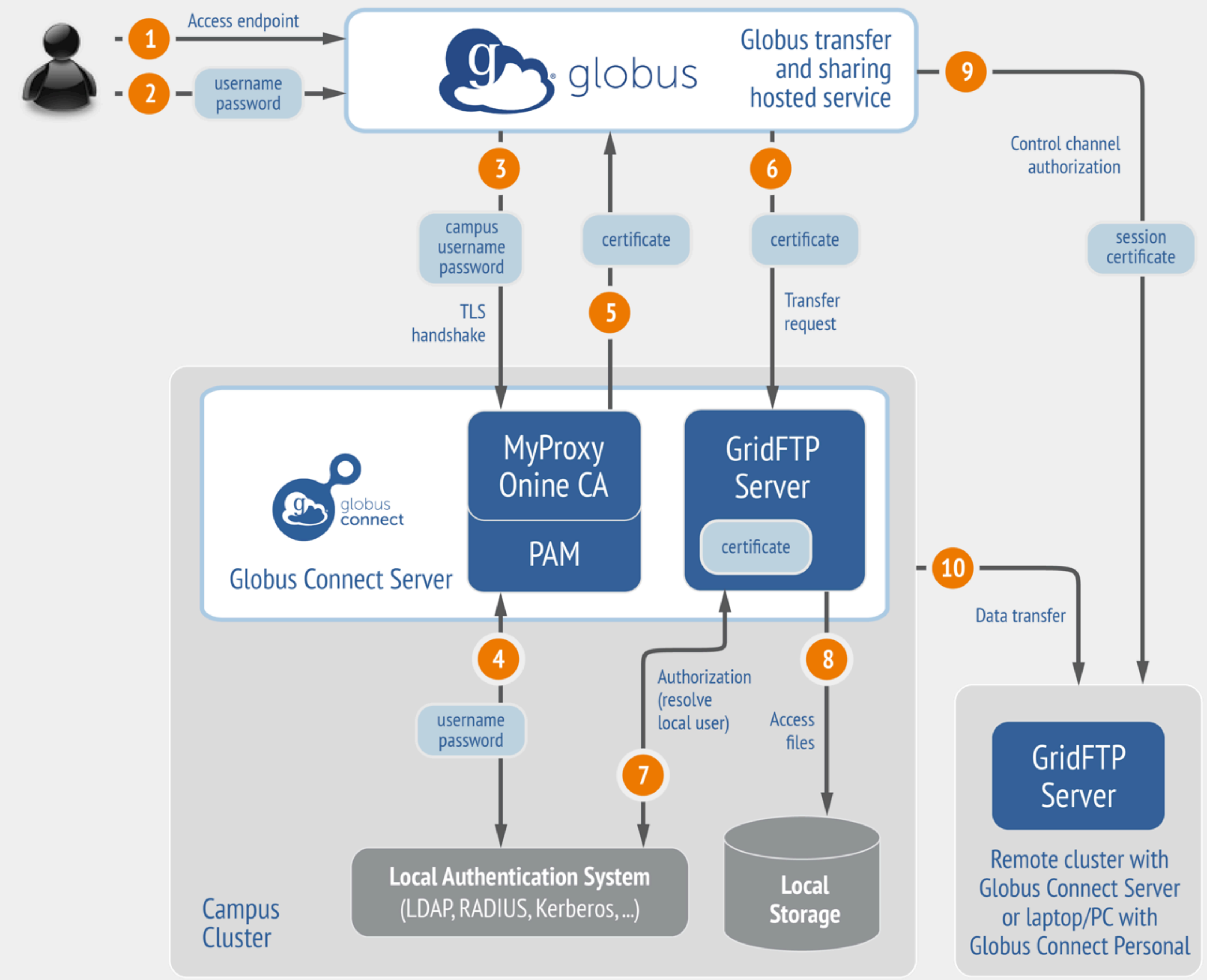
Folder	
Folder	
Folder	

625E1AAXX.1.1.fastq.gz	794 MB
625E1AAXX.1.1.fastq.gz.md5	57 b
625E1AAXX.1.2.fastq.gz	2.77 GB
625E1AAXX.1.2.fastq.gz.md5	57 b
625E1AAXX.2.2.fastq.gz	2.72 GB
625E1AAXX.2.2.fastq.gz.md5	57 b
EAS192_02_lane1.fastq.gz	6.59 GB
EAS192_02_lane1.fastq.gz.md5	77 b
EAS192_02_lane2.fastq.gz	6.55 GB
EAS192_02_lane2.fastq.gz.md5	77 b

3 **Access newly created endpoint**

4 **Transfer a file**

Endpoint activation using MyProxy



1 Access endpoint

2 username password

Globus transfer and sharing hosted service

3 campus username password

TLS handshake

5 certificate

6 certificate

Transfer request

9 Control channel authorization

session certificate

Globus Connect Server

MyProxy Onine CA

PAM

GridFTP Server

certificate

4 username password

7 Authorization (resolve local user)

8 Access files

10 Data transfer

Local Authentication System (LDAP, RADIUS, Kerberos, ...)

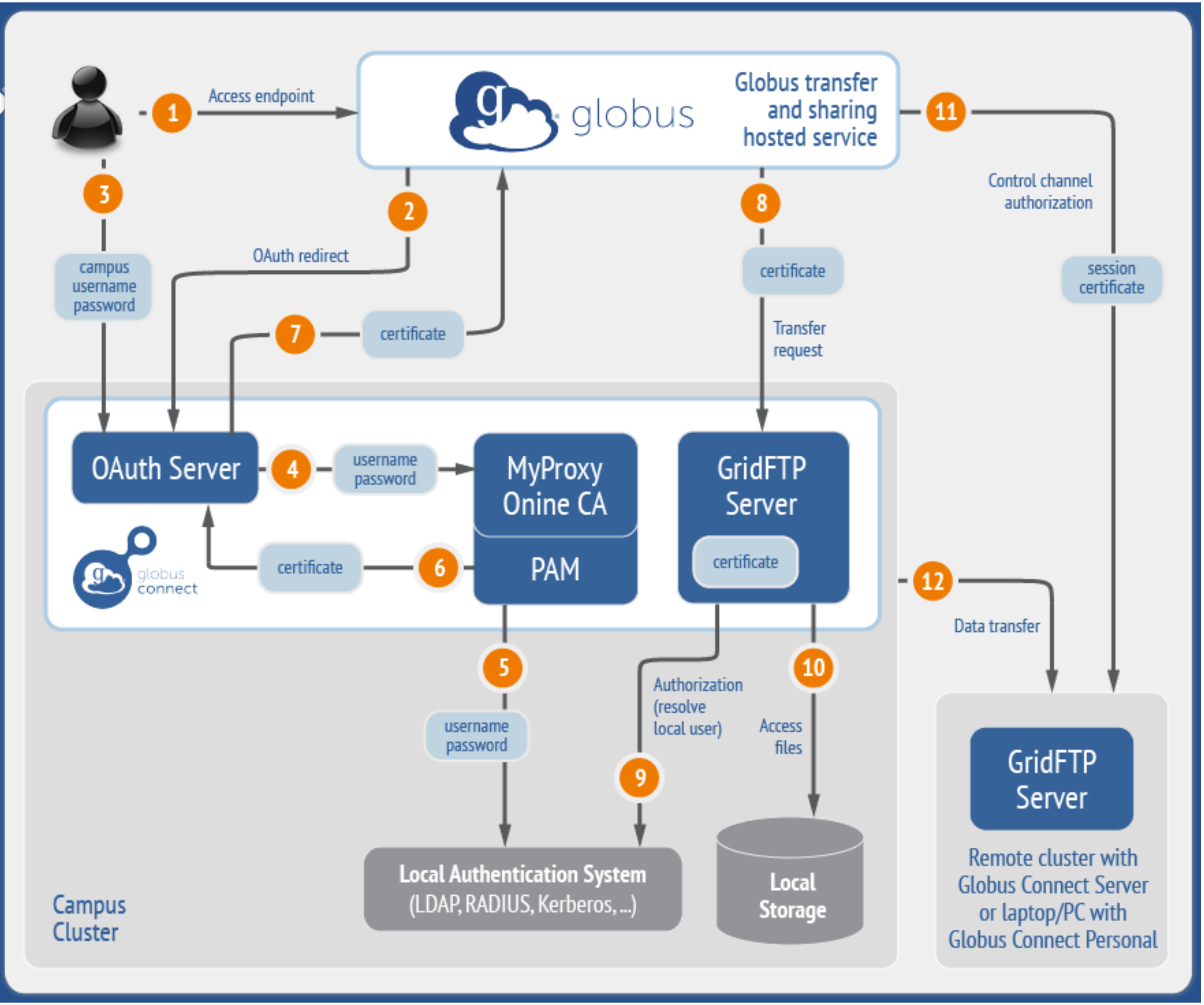
Local Storage

Campus Cluster

GridFTP Server

Remote cluster with Globus Connect Server or laptop/PC with Globus Connect Personal

Endpoint activation using MyProxy OAuth

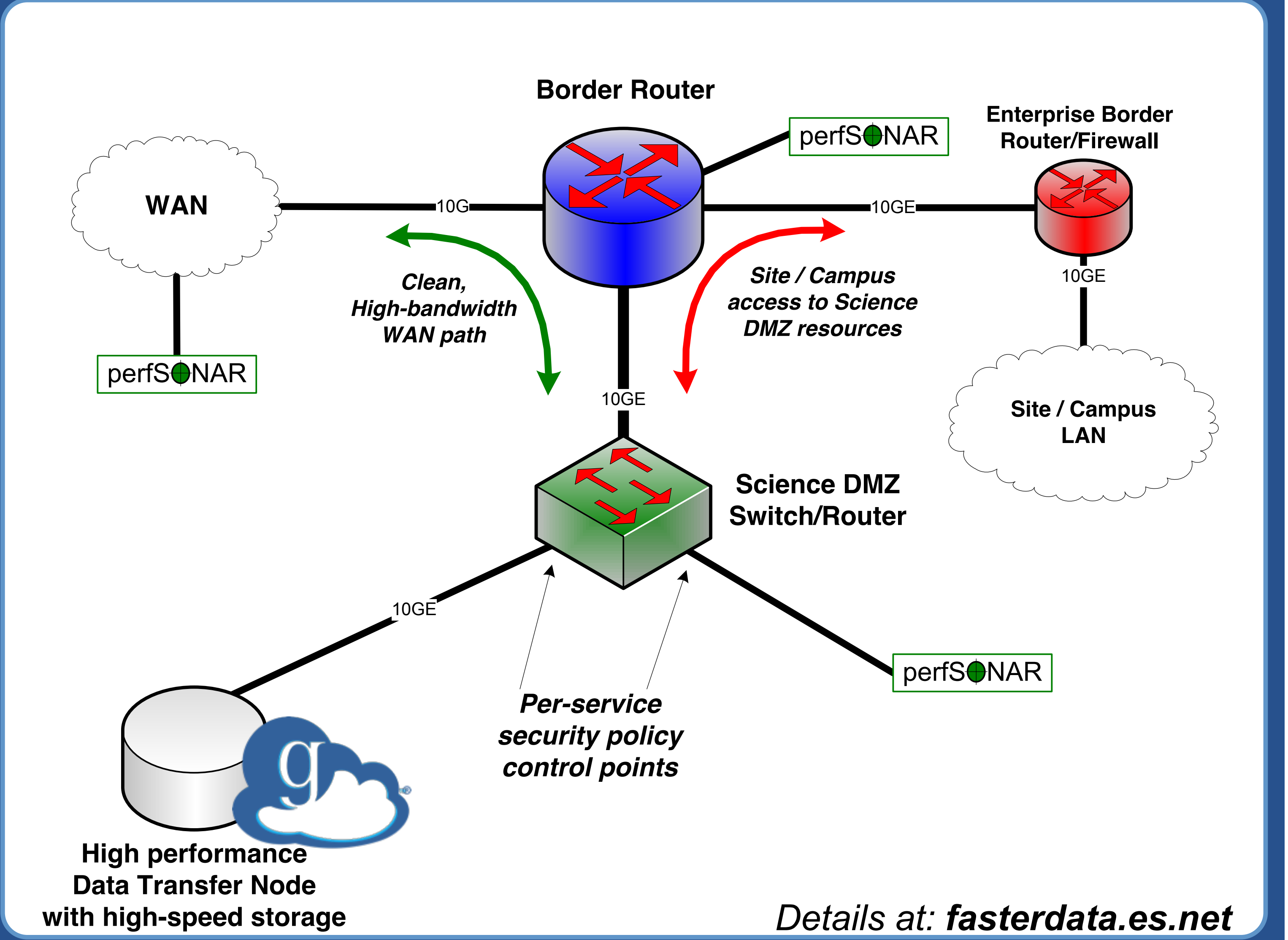


Common Configurations

- **Change endpoint name**
- **Customize filesystem access**
- **Enable sharing; set path restrictions**
- **Integrate with campus identity system**
- **Scale your campus deployment**
  - Data node(s)
  - Science DMZ

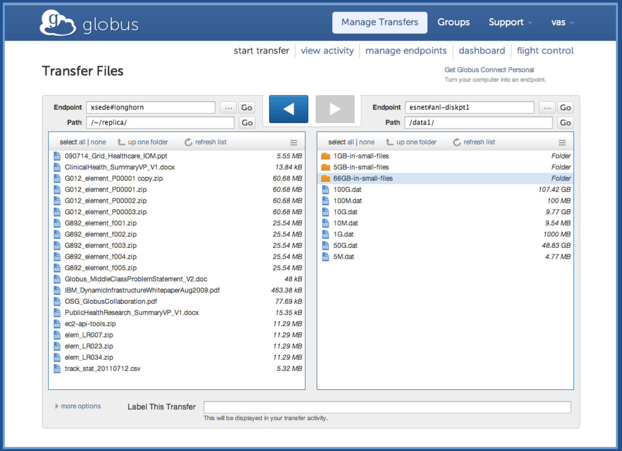
Deployment best practice

Science DMZ + Globus


Border Router
perfSONAR
Enterprise Border Router/Firewall
WAN
10G
10GE
Clean, High-bandwidth WAN path
Site / Campus access to Science DMZ resources
10GE
perfSONAR
10GE
Site / Campus LAN
Science DMZ Switch/Router
10GE
perfSONAR
Per-service security policy control points
High performance Data Transfer Node with high-speed storage

Details at: ***fasterdata.es.net***

Use(r)-appropriate interfaces



Transfer Files

Endpoint xsede#longhorn ... Go
Path /~/replica/ Go

090714_Grid_Healthcare_IOM.ppt	5.55 MB
ClinicalHealth_SummaryVP_V1.docx	13.84 kB
G012_element_P00001 copy.zip	60.68 MB
G012_element_P00001.zip	60.68 MB
G012_element_P00002.zip	60.68 MB
G012_element_P00003.zip	60.68 MB
G892_element_f001.zip	25.54 MB
G892_element_f002.zip	25.54 MB
G892_element_f003.zip	25.54 MB
G892_element_f004.zip	25.54 MB
G892_element_f005.zip	25.54 MB
Globus_MiddleClassProblemStatement_V2.doc	48 kB
IBM_DynamicInfrastructureWhitepaperAug2009.pdf	463.38 kB
OSG_GlobusCollaboration.pdf	77.69 kB
PublicHealthResearch_SummaryVP_V1.docx	15.35 kB
ec2-api-tools.zip	11.29 MB
elem_LR007.zip	11.29 MB
elem_LR023.zip	11.29 MB
elem_LR034.zip	11.29 MB
track_stat_20110712.csv	5.32 MB

Endpoint esnet#anl-diskpt1 ... Go
Path /data1/ Go

1GB-in-small-files	Folder
5GB-in-small-files	Folder
66GB-in-small-files	Folder
100G.dat	107.42 GB
100M.dat	100 MB
10G.dat	9.77 GB
10M.dat	9.54 MB
1G.dat	1000 MB
50G.dat	48.83 GB
5M.dat	4.77 MB

Label This Transfer

This will be displayed in your transfer activity.

Web

CLI

```
laptop:~ ssh vas@cli.globusonline.org
$ Welcome to globusonline.org, vas.  Type 'help' for help.
$ endpoint-modify vas#ebs --organization="University of Chicago”
$
```

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

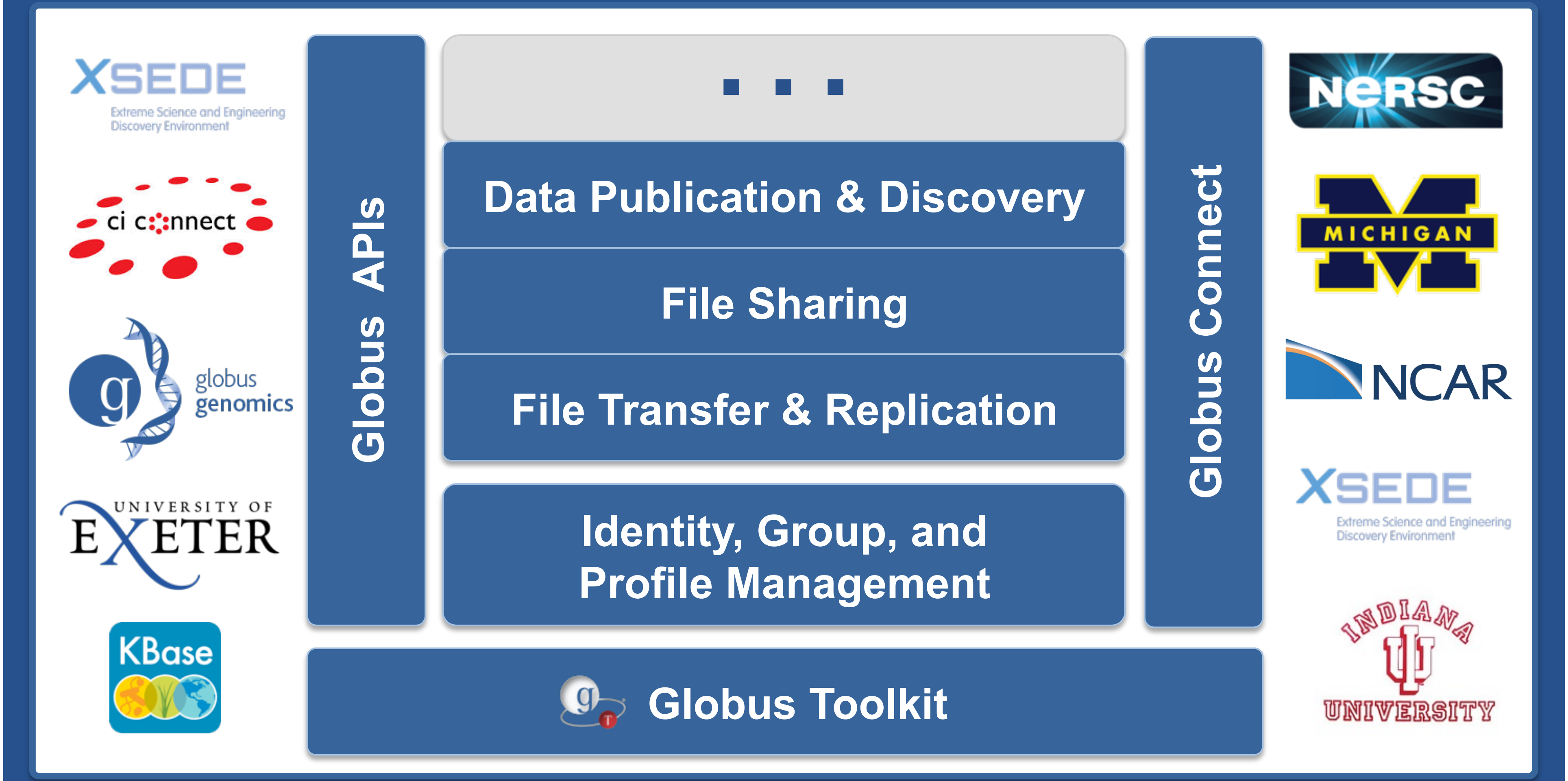
Rest API

Globus service

Quick look:

Globus Command Line Interface (CLI)

Globus Platform-as-a-Service



Globus APIs

■ ■ ■

Data Publication & Discovery

File Sharing

File Transfer & Replication

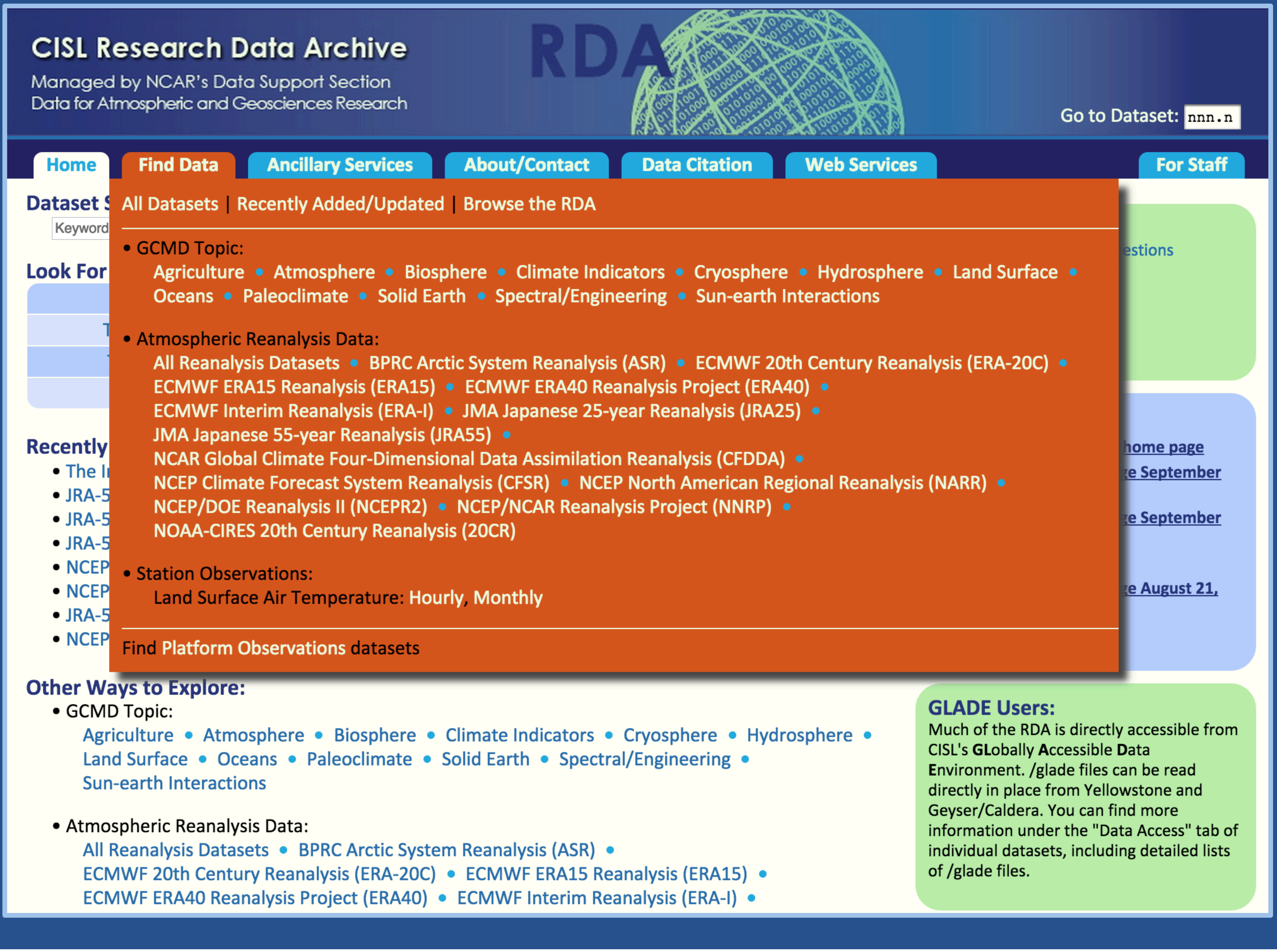
Identity, Group, and Profile Management

Globus Connect

Globus Toolkit

XSEDE
Extreme Science and Engineering Discovery Environment
ci connect
globus genomics
UNIVERSITY OF EXETER
KBase
NERSC
MICHIGAN
NCAR
XSEDE
Extreme Science and Engineering Discovery Environment
INDIANA UNIVERSITY

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: nnn.n
Home
Find Data
Ancillary Services
About/Contact
Data Citation
Web Services
For Staff

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

Dataset S

Keyword

Look For

Recently

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

estions

home page

e September

e September

e August 21,

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:

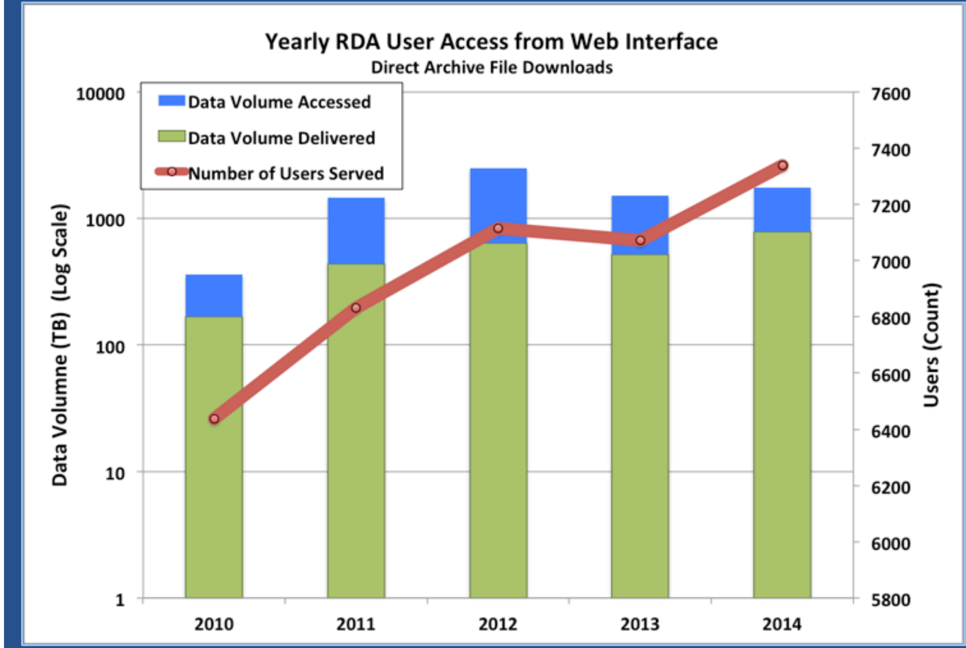
Much of the RDA is directly accessible from CISL's **GL**obally **A**ccessible **D**ata **E**nvironment. /glade files can be read directly in place from Yellowstone and Geyser/Caldera. You can find more information under the "Data Access" tab of individual datasets, including detailed lists of /glade files.

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)

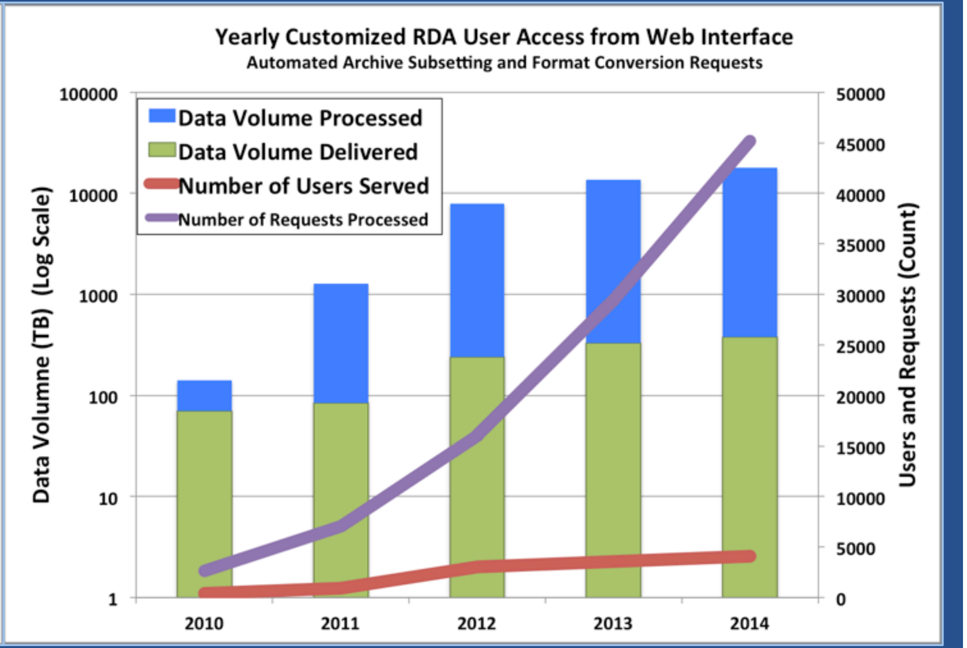
Courtesy of Thomas Cram, NCAR

RDA Usage



Yearly RDA User Access from Web Interface

Direct Archive File Downloads



Yearly Customized RDA User Access from Web Interface

Automated Archive Subsetting and Format Conversion Requests

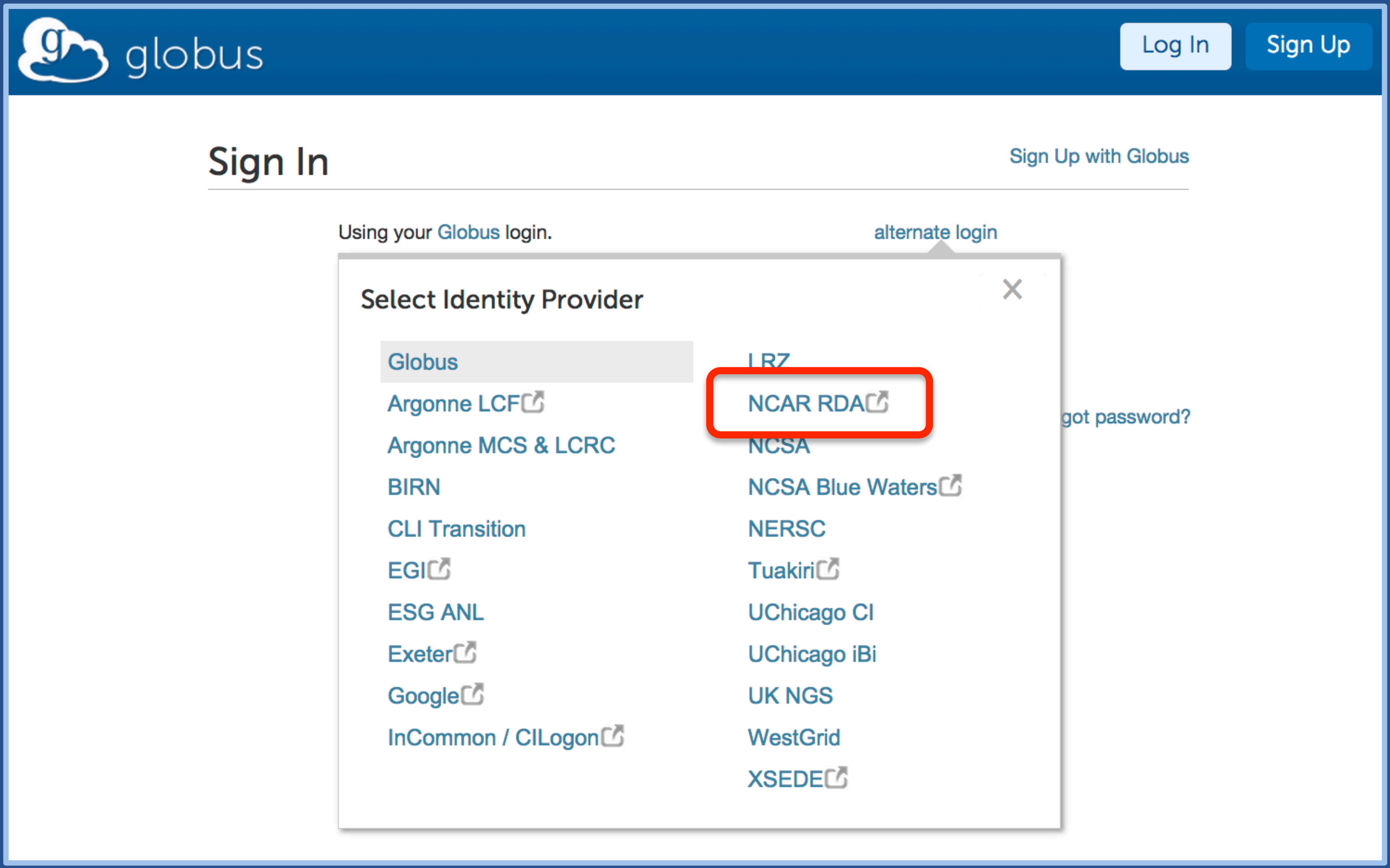
- **2014**
  - – 17+ PB virtual processing
  - – Web downloads: 7300 users, 750 TB served
  - – 45,000 custom orders, 4000 users, 380 TB served

Courtesy of Thomas Cram, NCAR

Globus @ RDA

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

RDA Alternate Identity login



globus

Log In

Sign Up

Sign In

Sign Up with Globus

Using your Globus login.

alternate login

Select Identity Provider

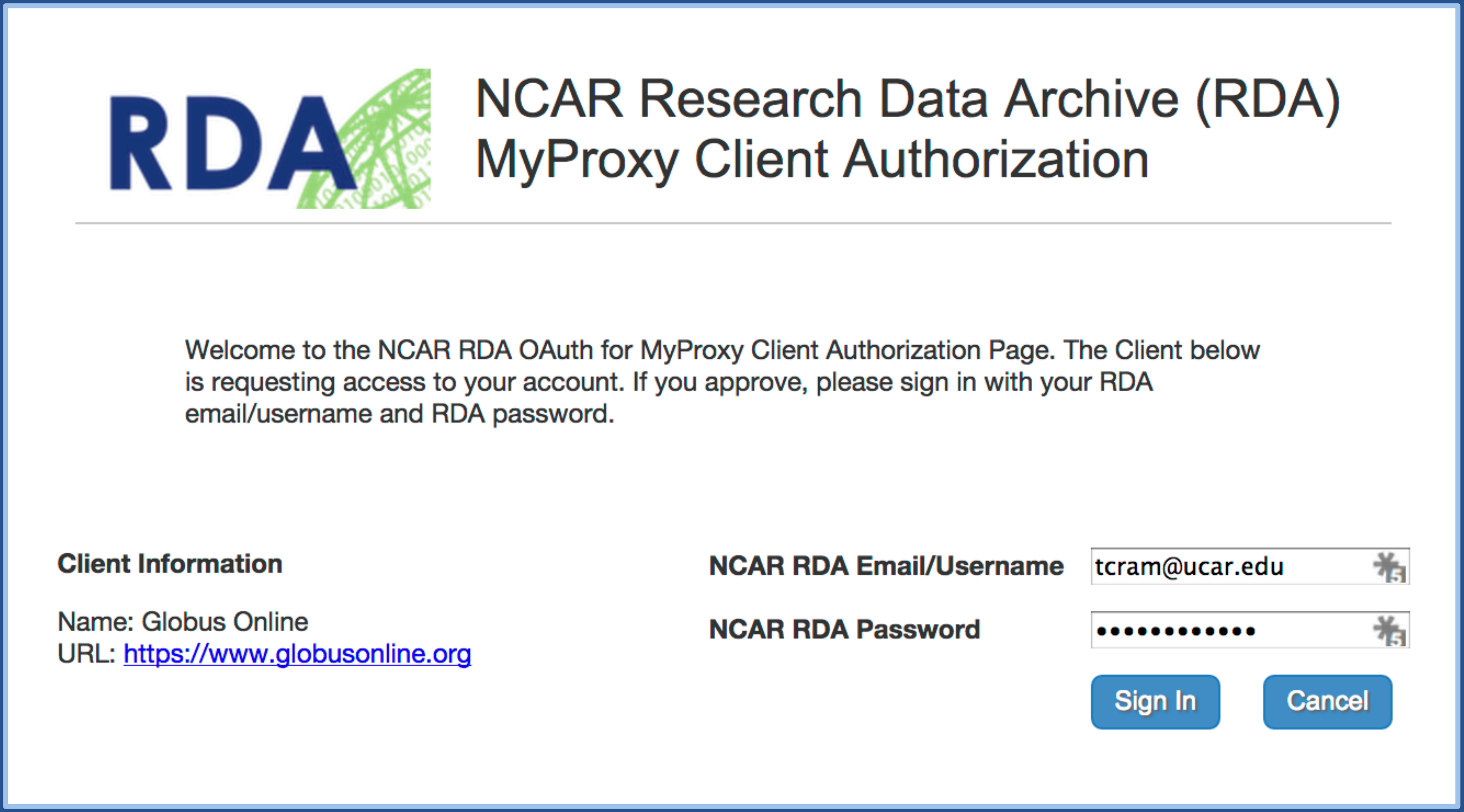
×

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

got password?

Courtesy of Thomas Cram, NCAR

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

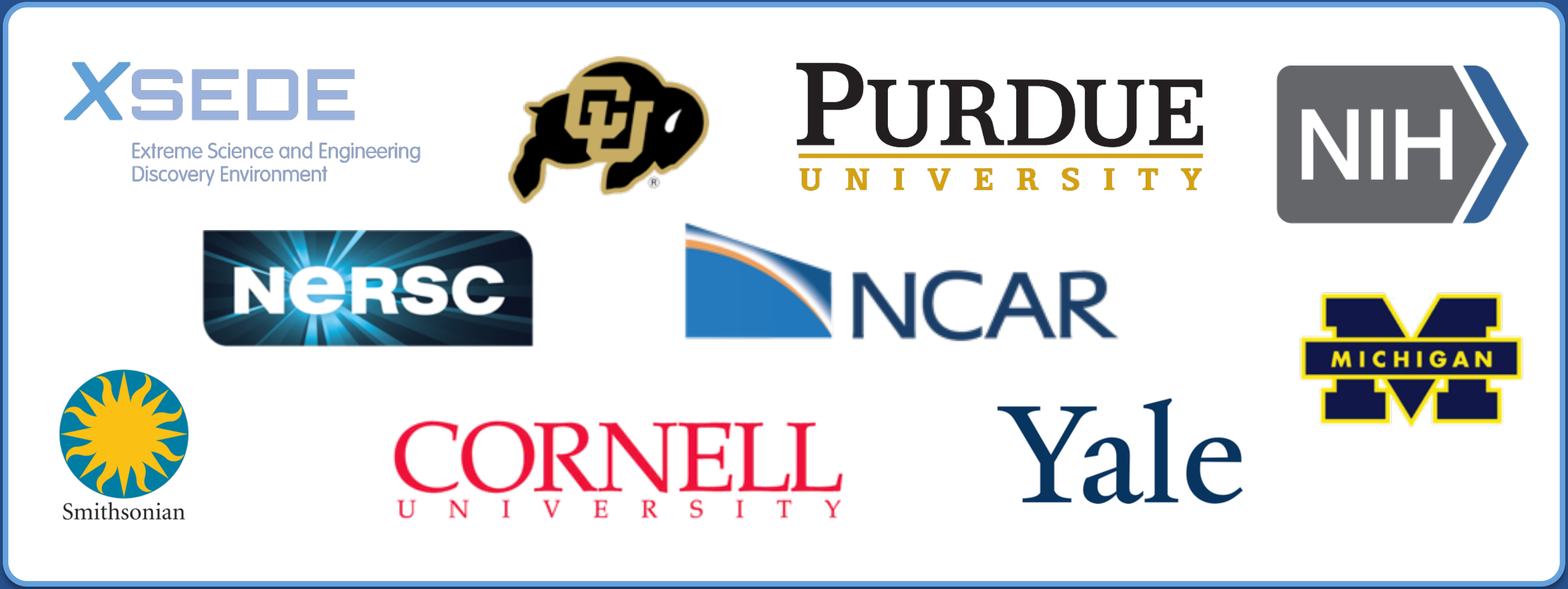
Name: Globus Online
URL: https://www.globusonline.org

NCAR RDA Email/Username tcram@ucar.edu

NCAR RDA Password •••••••••••••

Courtesy of Thomas Cram, NCAR

Some early Globus supporters



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**

Thank you