

Enhanced Research Data Management and Publication with Globus

Vas Vasiliadis
Jim Pruyne



THE UNIVERSITY OF
CHICAGO

Presented at OR2015
June 8, 2015





Presentations and other useful
information available at

globus.org/events/or2015/tutorial

bit.ly/or2015-globus



Thank you to our sponsors!



U.S. DEPARTMENT OF
ENERGY



THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY



powered by
amazon
web services



Agenda

- **Research data management scenarios and introduction to Globus**
- **Demonstrations and Exercises**
 - Accessing Globus and moving data
 - Sharing data and group management
 - Data publication and discovery
 - Creating collections
- **Campus deployment overview**
- **Globus today and tomorrow**



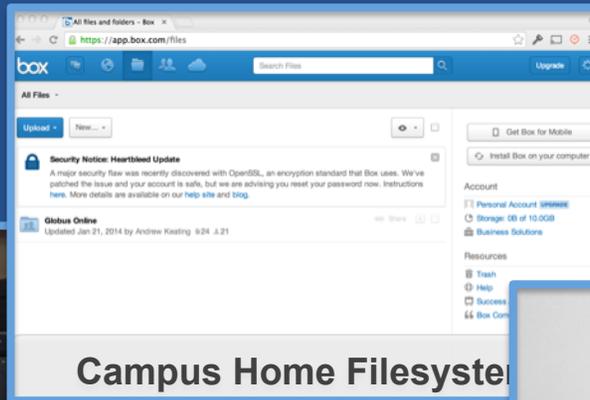
Research data management scenarios and challenges



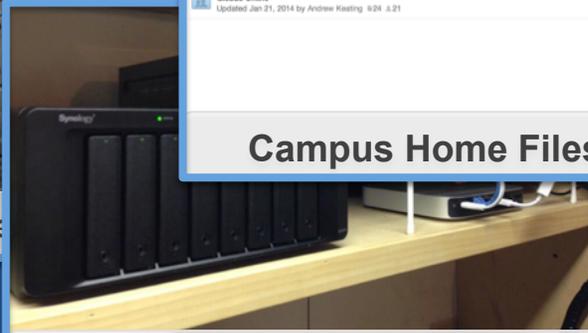
“I need to easily, quickly, & reliably move portions of my data to other locations.”



Research Computing HPC Cluster



Campus Home Filesystem



Lab Server



Personal Laptop



Desktop Workstation



XSEDE Resource



Public Cloud

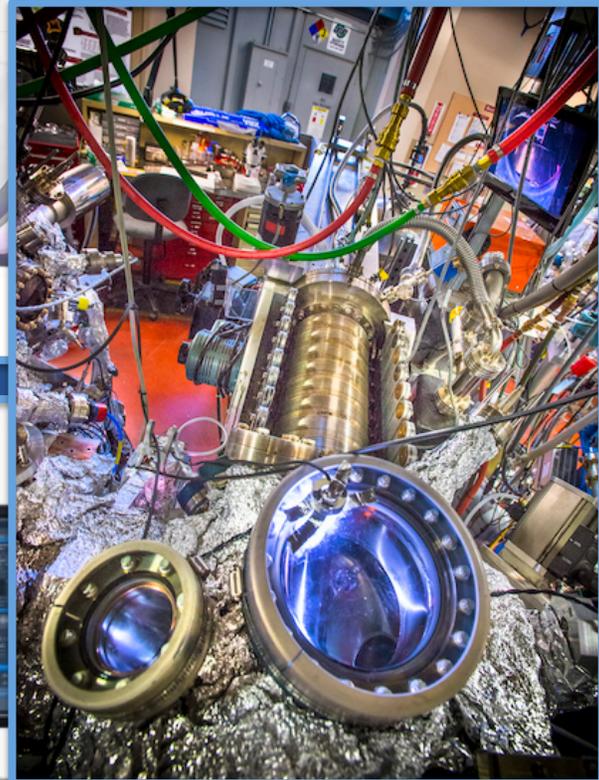


“I need to get data from a scientific instrument to my analysis system.”

MRI



Advanced Light Source



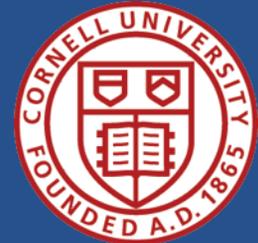
Next Gen Sequencer



Light Sheet Microscope



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





“I need to publish my data so that others can find it and use it.”

Reference
Dataset



Scholarly
Publication



Research
Community
Collaboration



Globus is...

Research data management...

...delivered via SaaS



Globus delivers...

Big data transfer, sharing,
publication, and discovery...

...directly from your own
storage systems



Our focus: User Experience

flickr ...for your photos

Google  ...for your office docs

NETFLIX ...for your entertainment

 globus ...for your research data

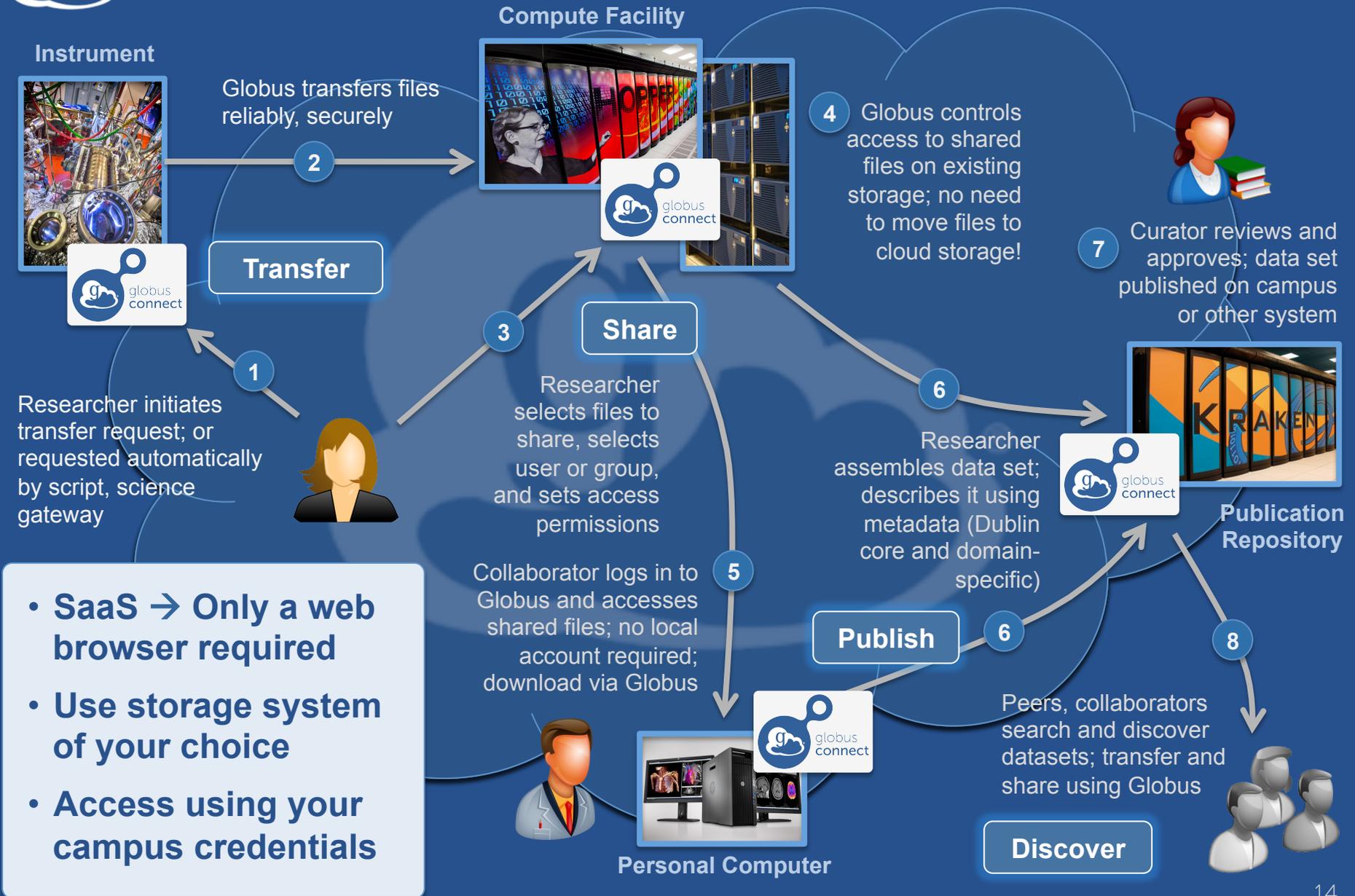


Globus is SaaS

- **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



Globus and the research data lifecycle





Demonstration: Accessing Globus and Moving Data



Exercise 1: 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 esnet#... endpoint to
your laptop**



Demonstration: Sharing Data



Exercise 2: 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**



Demonstration: 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

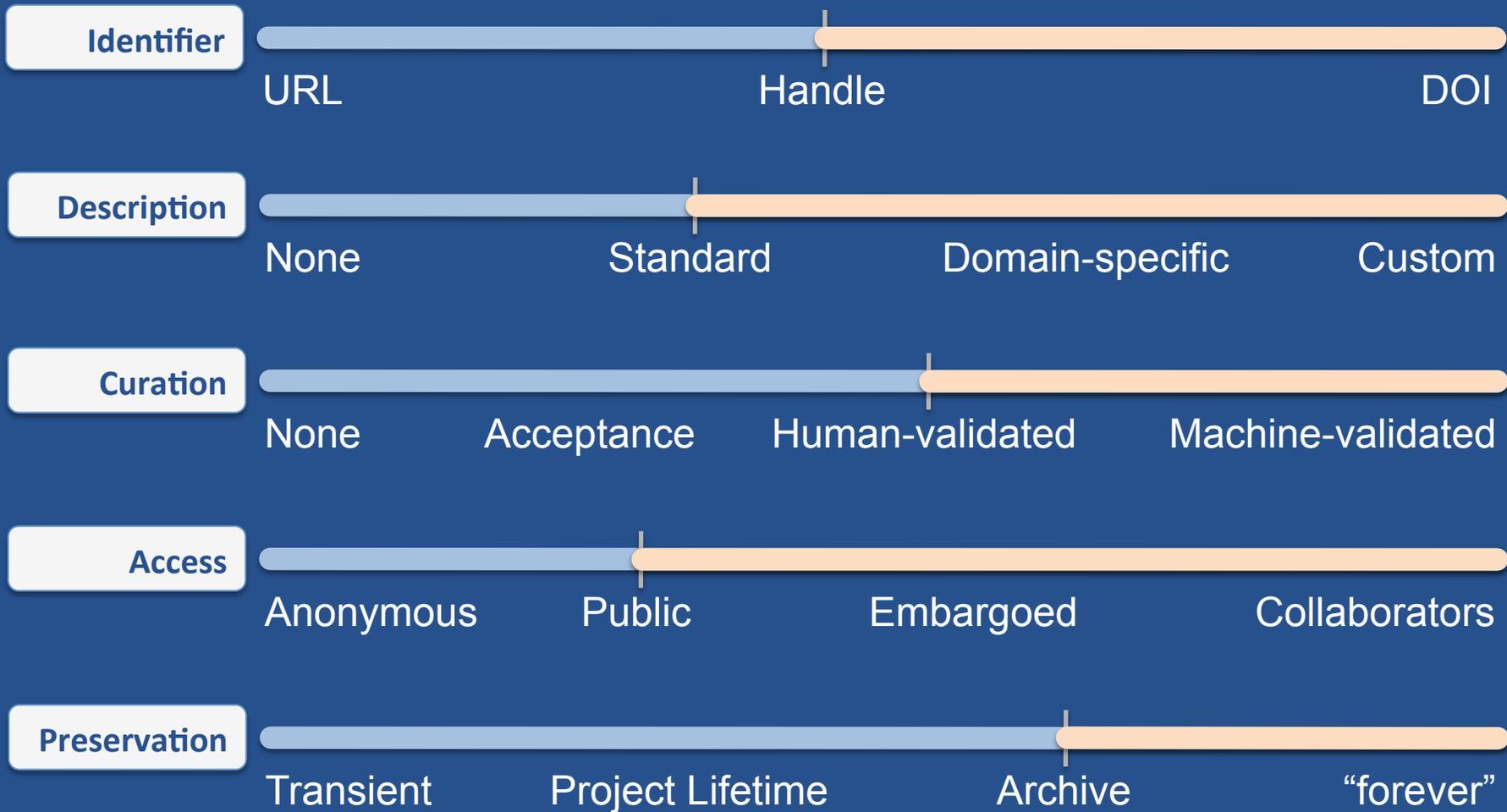
5. Confirm your neighbor can access your shared endpoint



“Appropriate” preservation



Globus data publication framework



Publish

Submit: Describe this Dataset

Title *

Authors *

Publication Year *

Publisher *

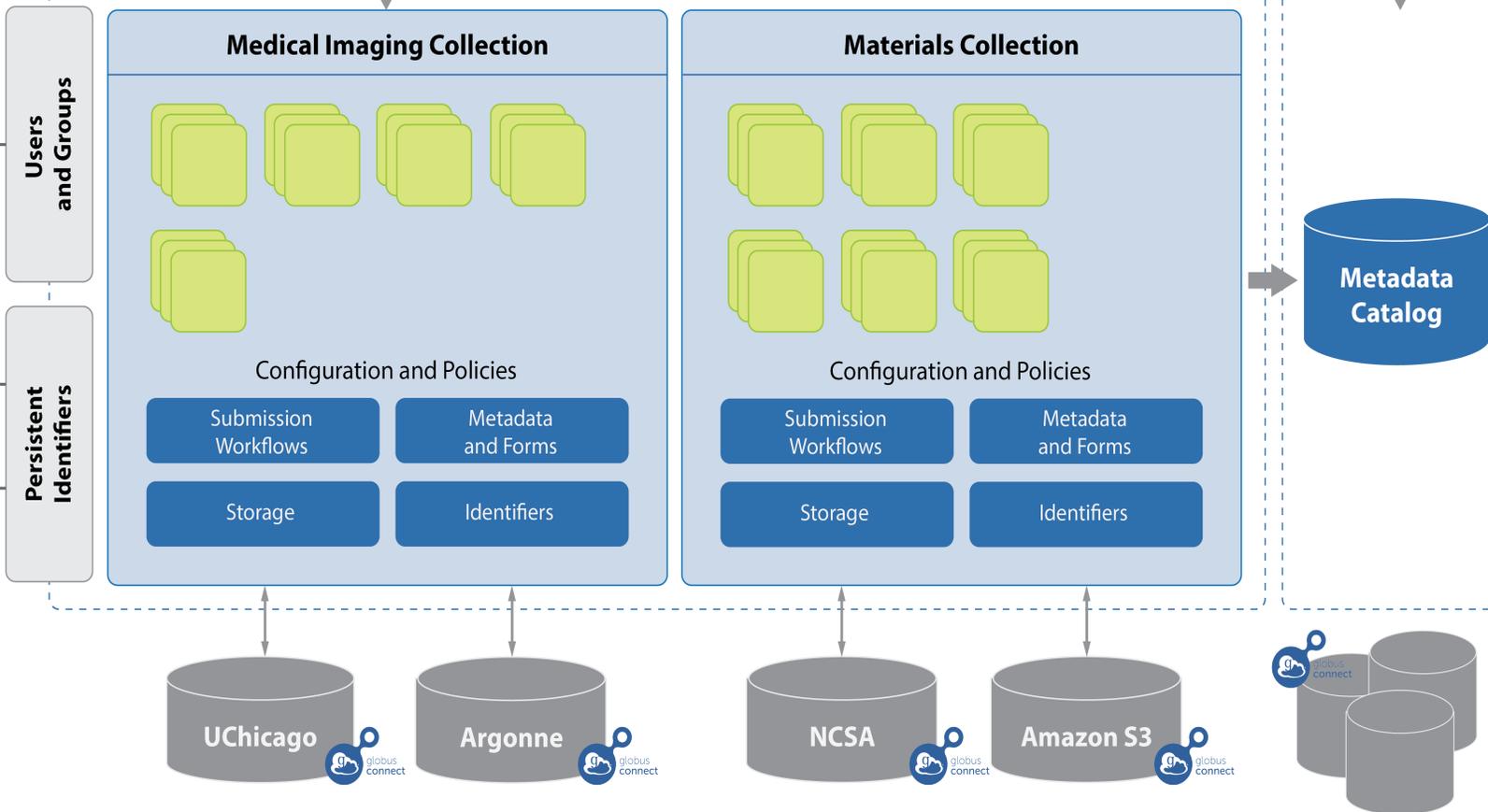


Discover

Issue Date	Title	Author(s)
31-Jul-2014	Sample Data	UCSD, MIT, Max Planck Society
1990	Contingency Tables	UCSD, MIT, Max Planck Society
1990	Thermal Conductivity	UCSD, MIT, Max Planck Society
1990	Thermal Conductivity	UCSD, MIT, Max Planck Society
1990	Thermal Conductivity	UCSD, MIT, Max Planck Society

Globus Authentication

Globus Data Publication





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=1,Type=Float,Desc
```

Processing metadata...

- Pipeline description
- Tool parameters
- Exec environment

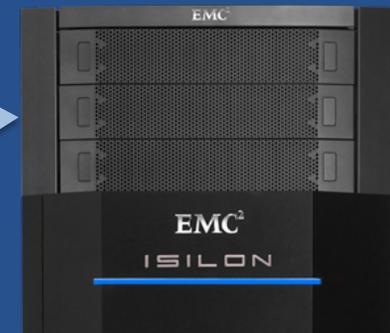
Automated
curation

- Machine validated
- Exception review

Identify...

- URL

Moderate
durability/cost





Downstream analysis



XSEDE
Extreme Science and Engineering
Discovery Environment

Jetstream

EC2

NERSC

Optional metadata...

- “Implicit” metadata
- Description through organization

Team review

- Any collaborator may approve

Identify...

- Globus share

Widely accessible stores

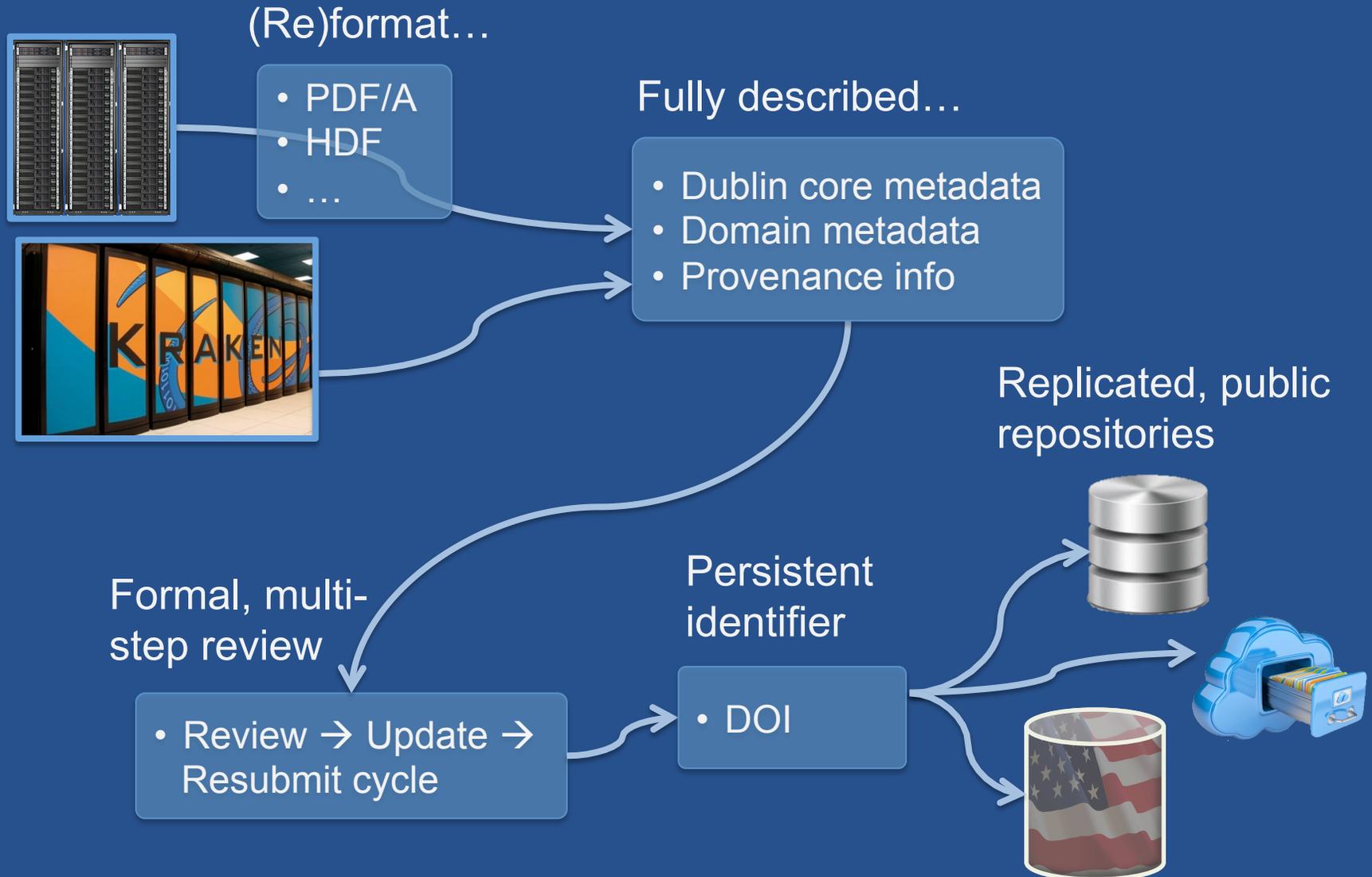
amazon web services™ **S3**

Jetstream

red cloud



Peer reviewed paper





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”



Demonstration: Data Publication and Discovery



Exercise 4: Publish a dataset

1. Go to trial.publish.globus.org
2. Log in, click “Start a New Submission”
3. Select the “Globus Demo Collection”
4. Agree to the license
5. Enter the required metadata
6. Assemble data set from the `esnet#...` Endpoints (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



Example use cases



Repository planning



NEW YORK UNIVERSITY

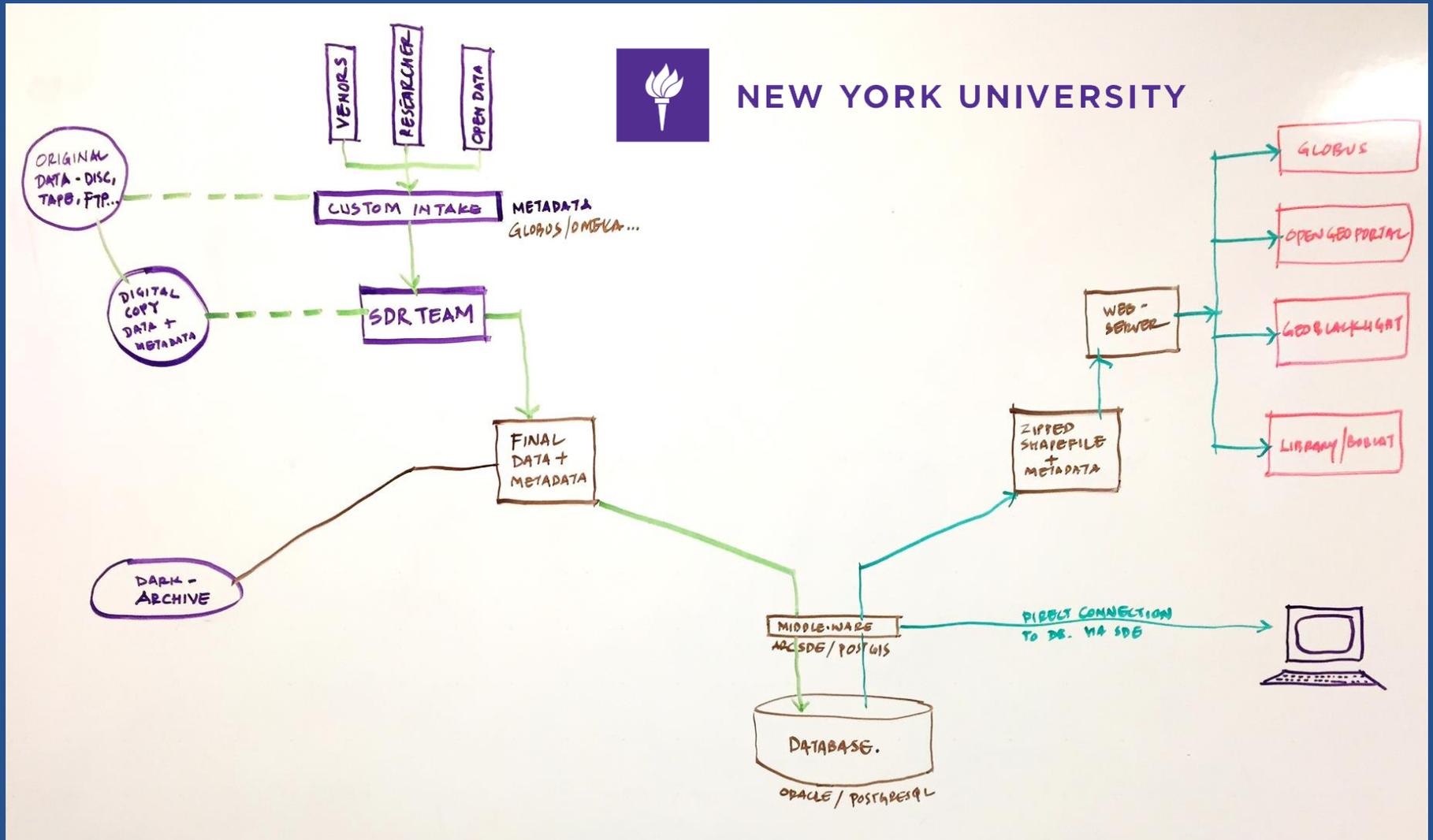
Mix-Model: Technology meets Pedagogy

- **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



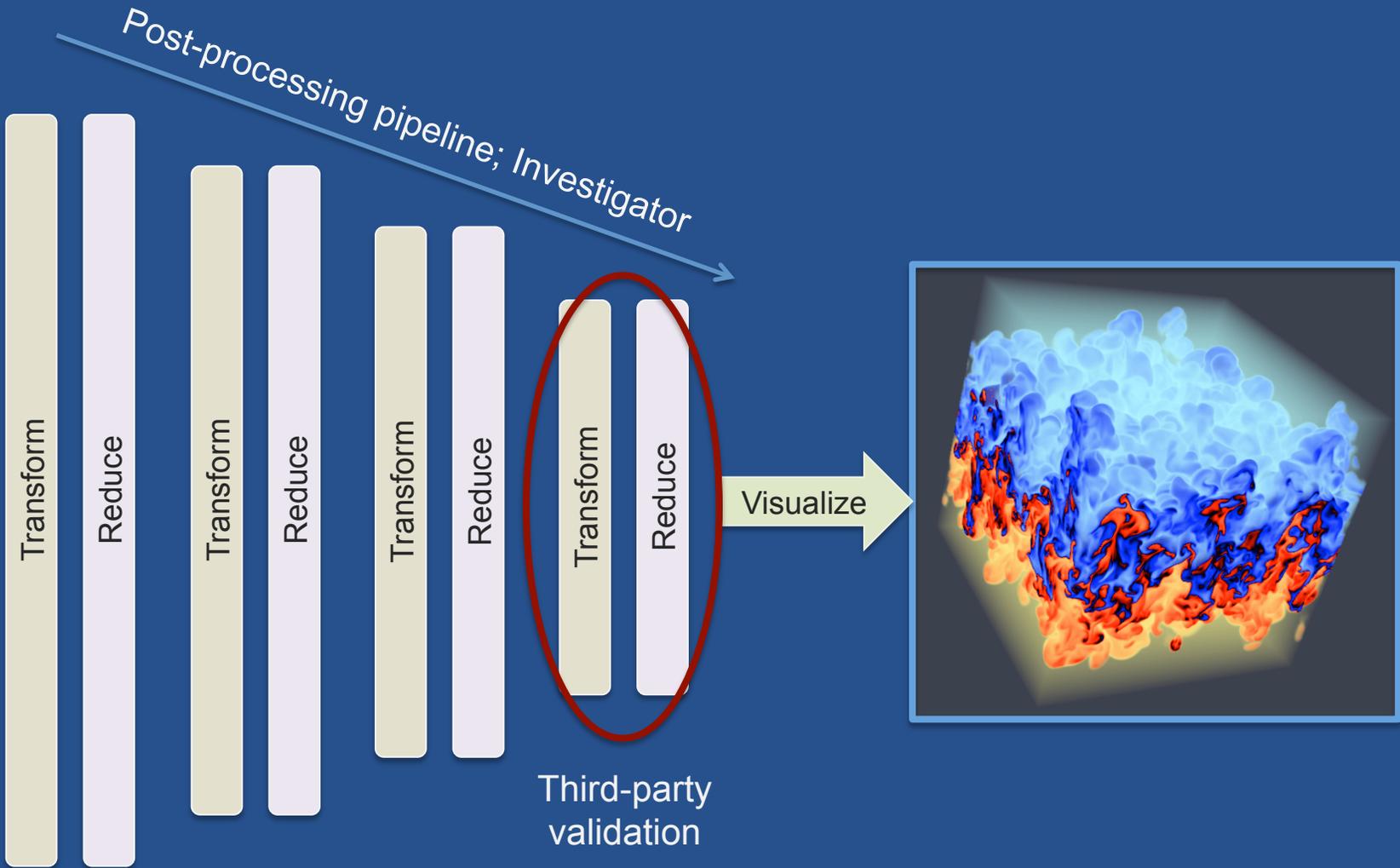
Repository planning



Source: H. Mistry, New York University



Reproducibility



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



Reproducibility

nbviewer FAQ IPython Jupyter



nek-workflow / Demo.ipynb

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/"
```



Demonstration: Collection Configuration



Exercise 5: Create a collection

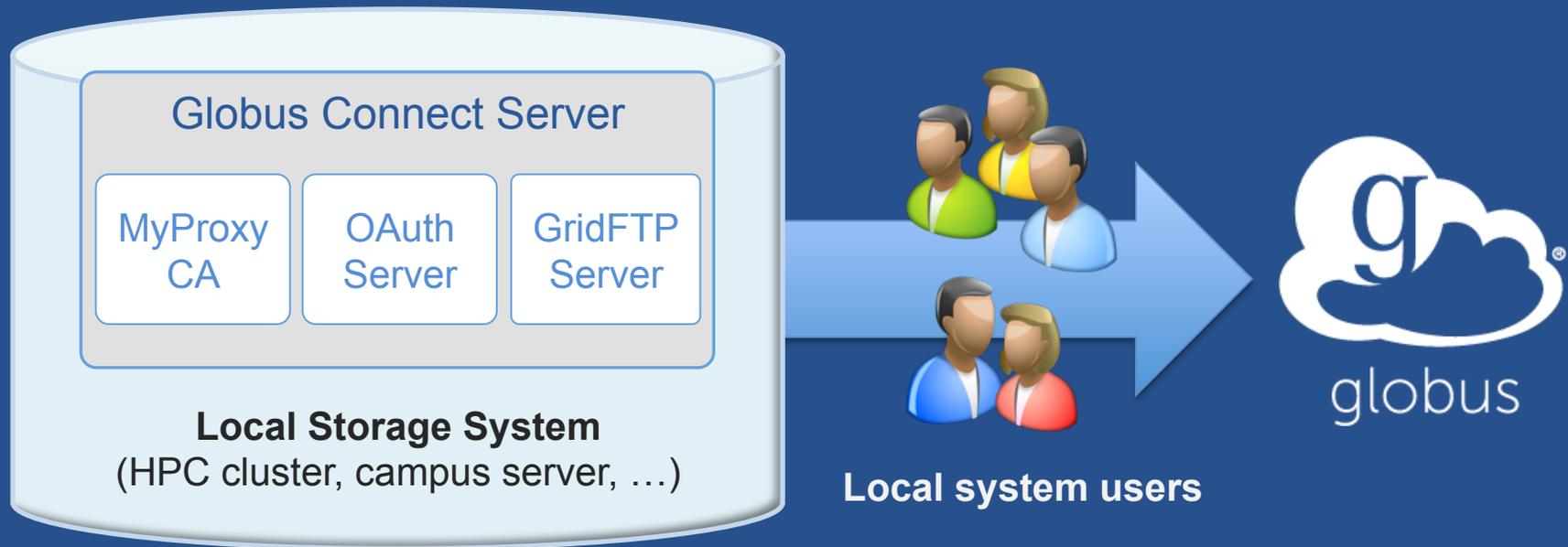
- 1. Create a new collection**
- 2. Enter metadata**
 - 1. Name and submission license (required)**
 - 2. Description (optional)**
- 3. Enter the endpoint for collection storage: globuspublish#or2015-tutorial**
- 4. Enter a prefix of your choice**
- 5. Identifier: Select OR2015: bit.ly**
- 6. Leave all other fields at default values**
- 7. Select Curation Group (Tutorial Users)**
- 8. Submit a dataset for publication into your new collection**
- 9. Review and approve your neighbor's submission**



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

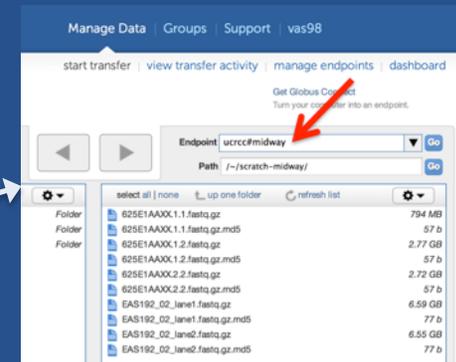


1 Install Globus Connect Server

- Access server as user "clusteradmin"
- Update repo
- Install package
- Setup Globus Connect Server



2 Log into Globus (using your Globus username)



3 Access the newly created endpoint (as user 'researcher')

4 Transfer a file



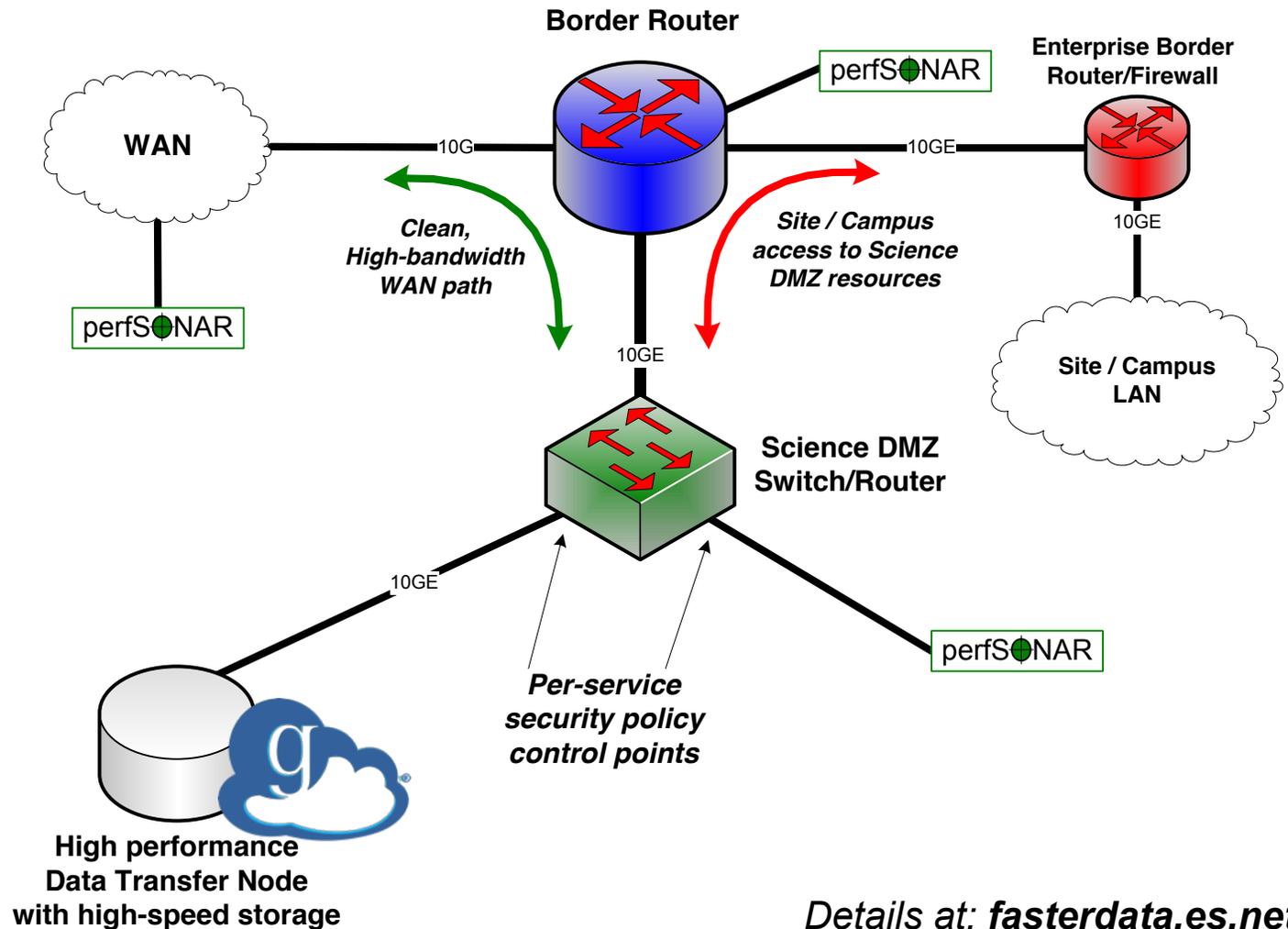
Common Configurations

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



Typical deployment

Science
DMZ
+
Globus





Demonstration: Globus Command Line Interface (CLI)



Globus: today and tomorrow



Globus today...

~ 100PB moved

>10,000 endpoints

>300 active users/day



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**

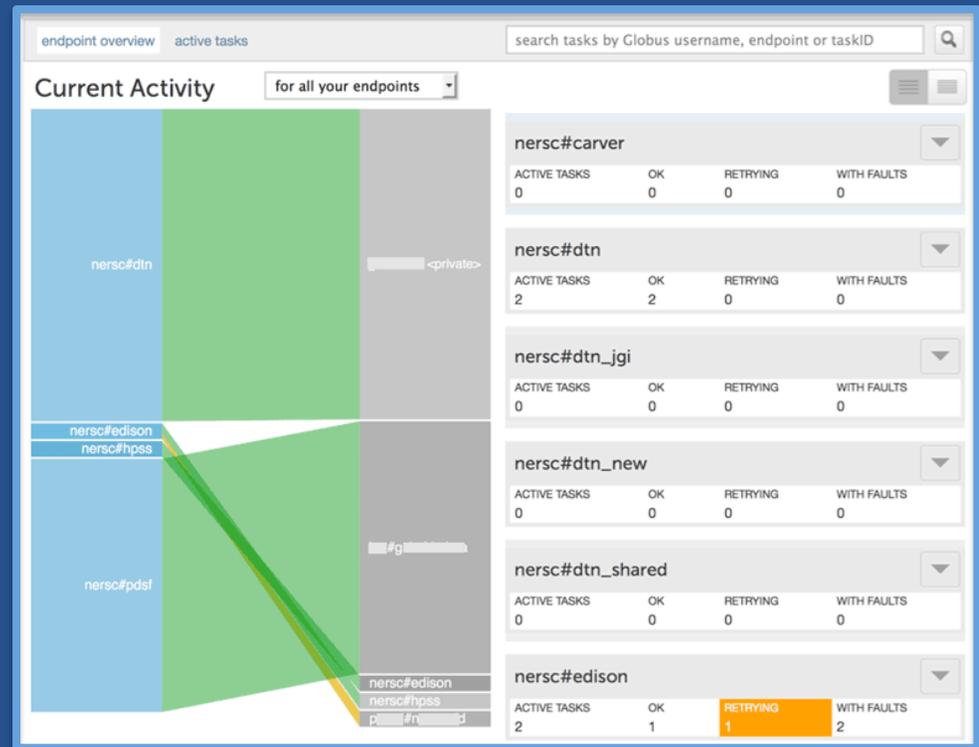
- Shared endpoints
- Data publication
- Amazon S3 endpoints
- Management console
- Usage reporting
- Priority support
- Application integration

- **Branded Web Site**

- **Alternate Identity Provider (InCommon is standard)**

- **Mass Storage System optimization**

globus.org/provider-plans





Demonstration: Globus management console



Globus Platform-as-a-Service





Some early Globus supporters

XSEDE

Extreme Science and Engineering
Discovery Environment



**Carnegie
Mellon
University**

**MICHIGAN STATE
UNIVERSITY**



Te Whare Wānanga o Tāmaki Makaurau



Information Sciences Institute



EMORY



**CORNELL
UNIVERSITY**



Ole Miss



THE UNIVERSITY OF
CHICAGO



NEW YORK UNIVERSITY





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