Globus: Simplifying Research Data Management via Software-as-a-Service

Vas Vasiliadis
University of Chicago, Argonne National Laboratory
vas@uchicago.edu

globus

4th Workshop on Promoting Open Science Data, March 1, 2017

My background

- 2010 – Present
  - Computation Institute, University of Chicago and Argonne National Laboratory
  - Masters Program in Computer Science: Lecturer in Cloud Computing and Product Management
  - Globus : Chief Customer Officer
- Prior Experience
  - Innovation consulting
  - Large scale systems integration
  - Real-time data distribution systems development
- Education
  - MBA, University of Michigan, Ann Arbor, MI, USA
  - BSEE, University of the Witwatersrand, South Africa

Agenda

- **Motivation**
- **Globus Software-as-a-Service**
- **Globus Platform-as-a-Service**
- **Building the research app ecosystem**
- **A word on sustainability**

Research data management

Index?

How do we...

...move?

...share?

...discover?

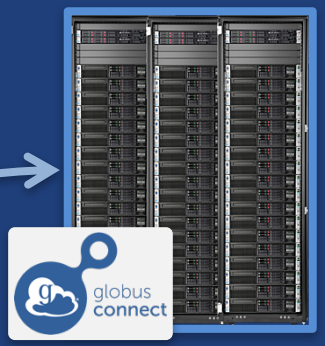
...reproduce?

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 cloud resources

XSEDE
Extreme Science and Engineering Discovery Environment

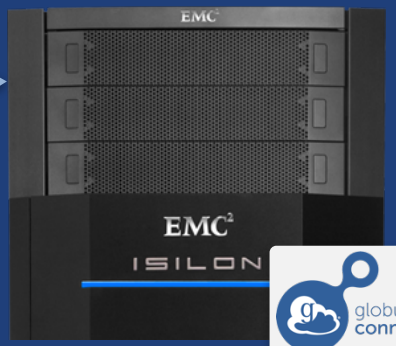
EC2

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

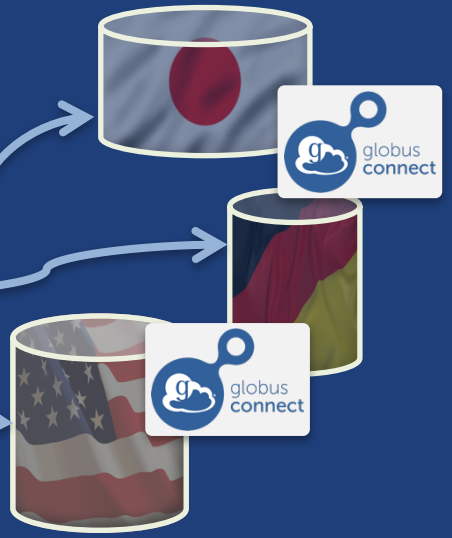
(Re)format…

- PDF/A
- HDF
- …

Describe…

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

Replicated, public Repositories (e.g. WDS)



Multi-step curation workflow

- Review → Update → Resubmit cycle

Persistent identifier

- DOI

Globus enables...

Collaboration

…within and beyond campus boundaries

Globus delivers…

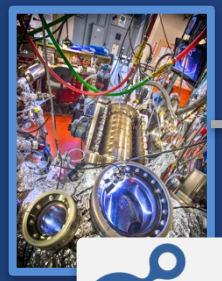
Big data transfer, sharing, publication, and discovery…

…directly from your own storage systems…

…via software-as-a-service

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

Transfer

Share

1

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

3

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

6

Researcher assembles data set; describes it using metadata (Datacite & domain-specific)

Publication Repository

5

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

Publish

6

8

- Access via Web browser
- Use any storage system
- Use existing identity

Personal Computer

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

Discover

Why use Globus?

- **Ease of use: consistent UI across systems**
- **“Fire-and-forget” file transfer**
- **No/low-overhead external collaboration**
- **Secure access w/multi-tier security model**
- **Maximized WAN throughput**
- **Rapid deployment w/standard packages**
- **Highly automatable: CLI, RESTful API**

Focus on User Experience

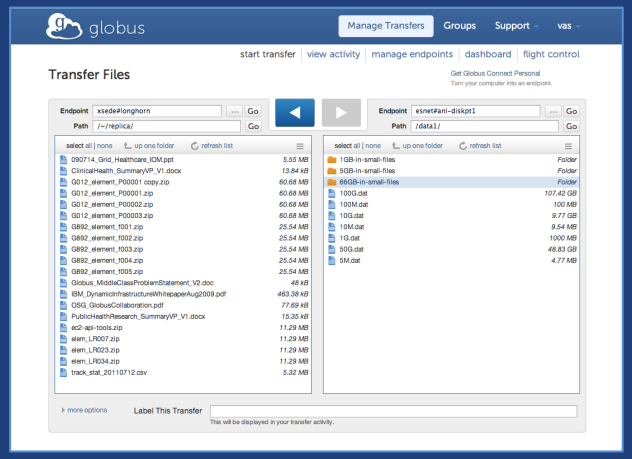
...for your photos

...for your office docs

...for your entertainment

...for your research data

Use(r)-appropriate interfaces



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"
$
```

Globus service

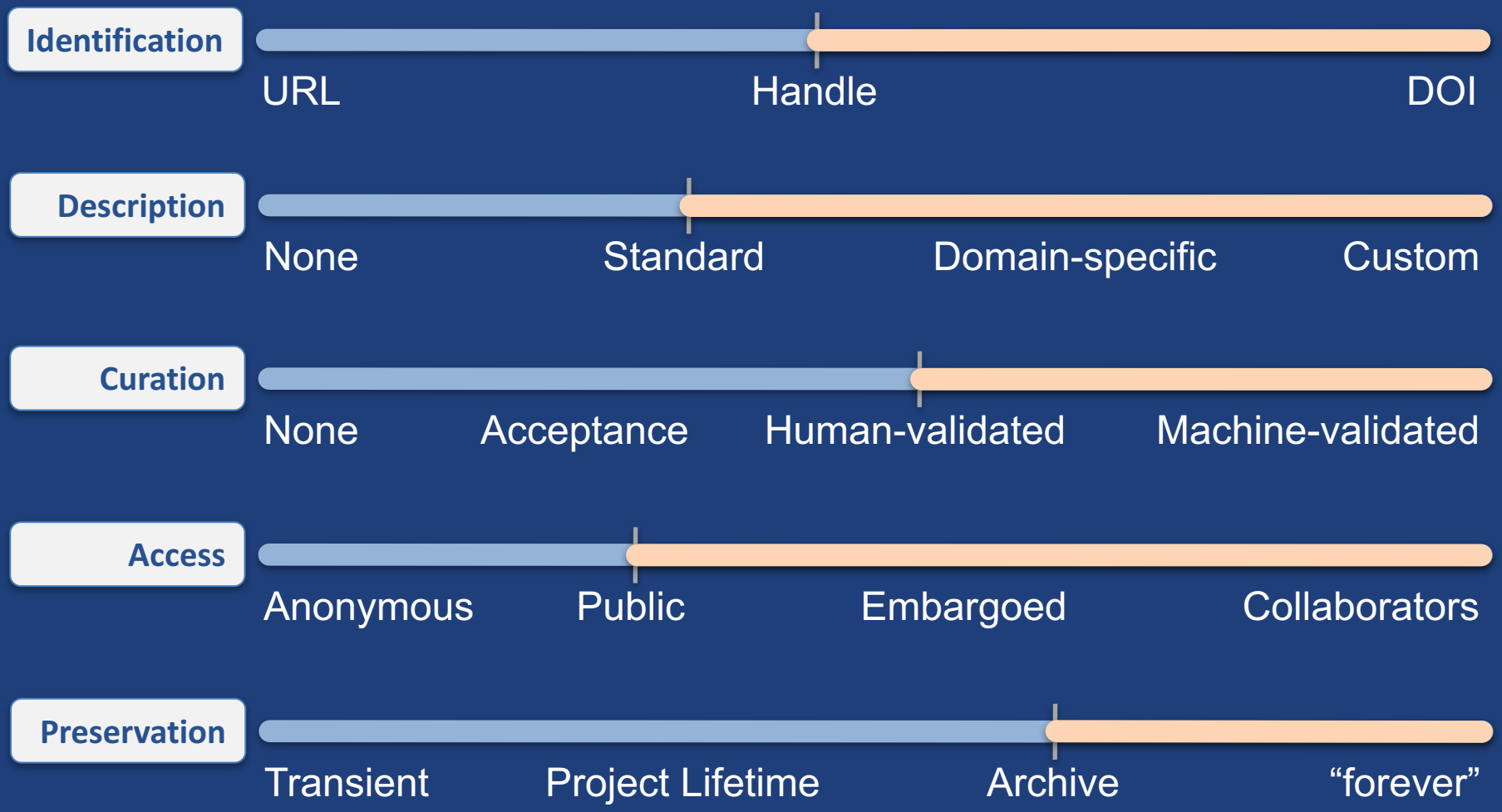
```
GET /endpoint/go%23ep1
PUT /endpoint/vas#my_endpt
200 OK
X-Transfer-API-Version: 0.10
Content-Type: application/json
…
```

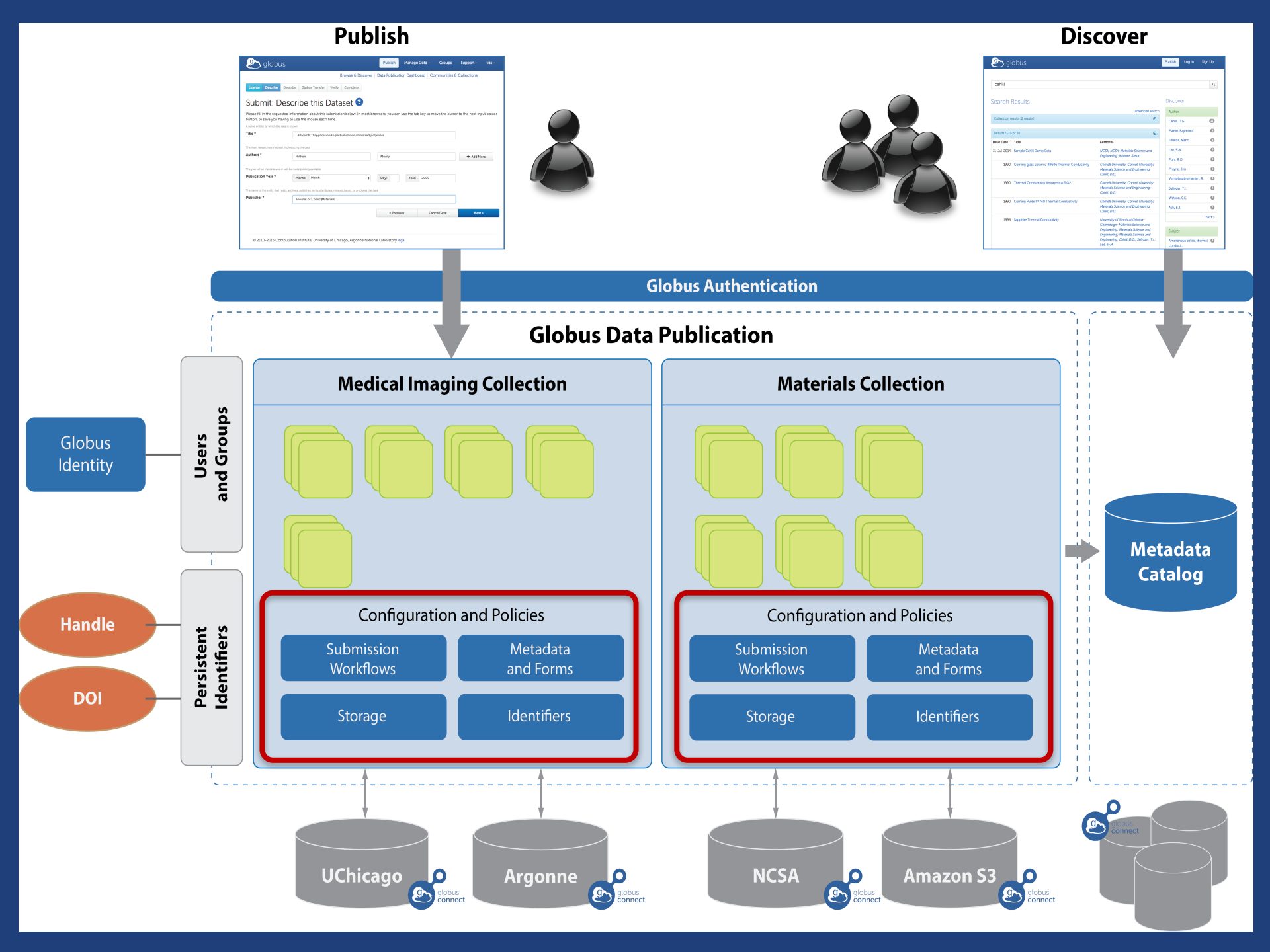
Rest API

Demonstration

- Identity Federation
- Transfer and Sharing

Globus data publication framework




Publish
Discover
Globus Authentication
Globus Data Publication
Medical Imaging Collection
Materials Collection
Configuration and Policies
Submission Workflows
Metadata and Forms
Storage
Identifiers
Users and Groups
Persistent Identifiers
Globus Identity
Handle
DOI
Metadata Catalog
UChicago
Argonne
NCSA
Amazon S3
globus connect

Demonstration
- Data Publication

Globus Connect…

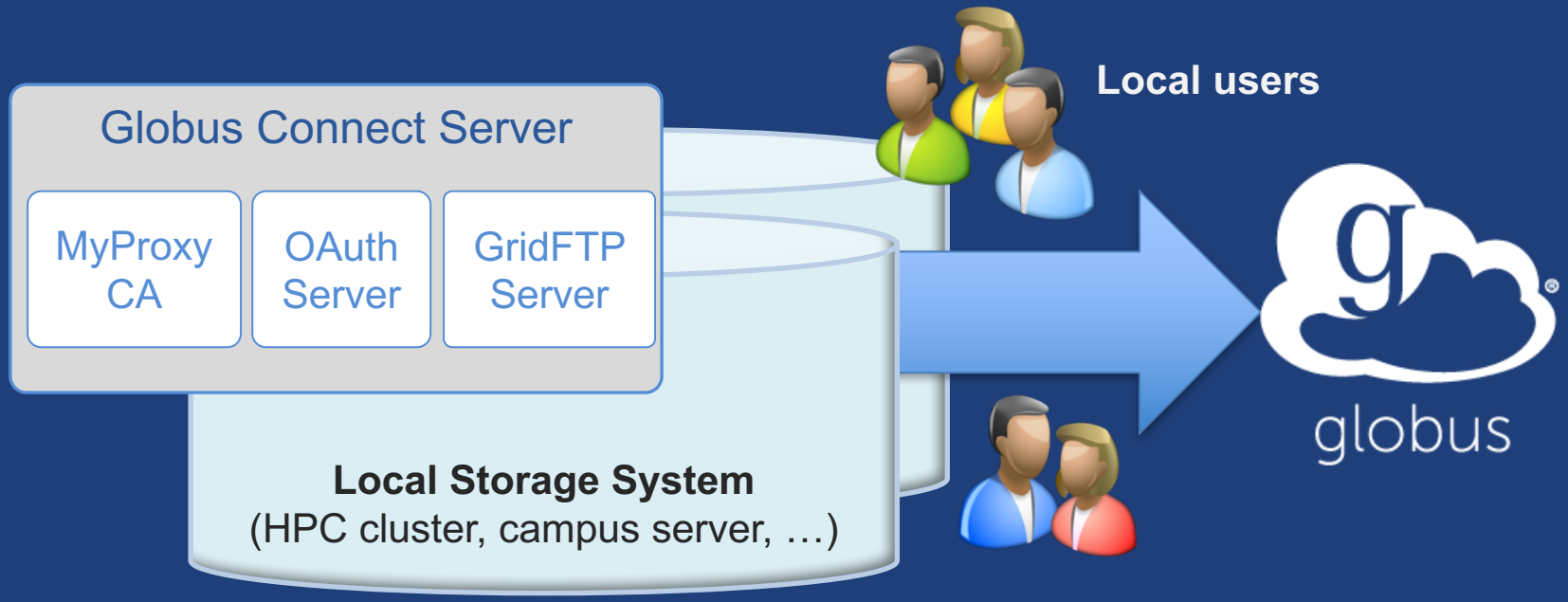
Makes your storage system a Globus endpoint

Globus Connect Personal



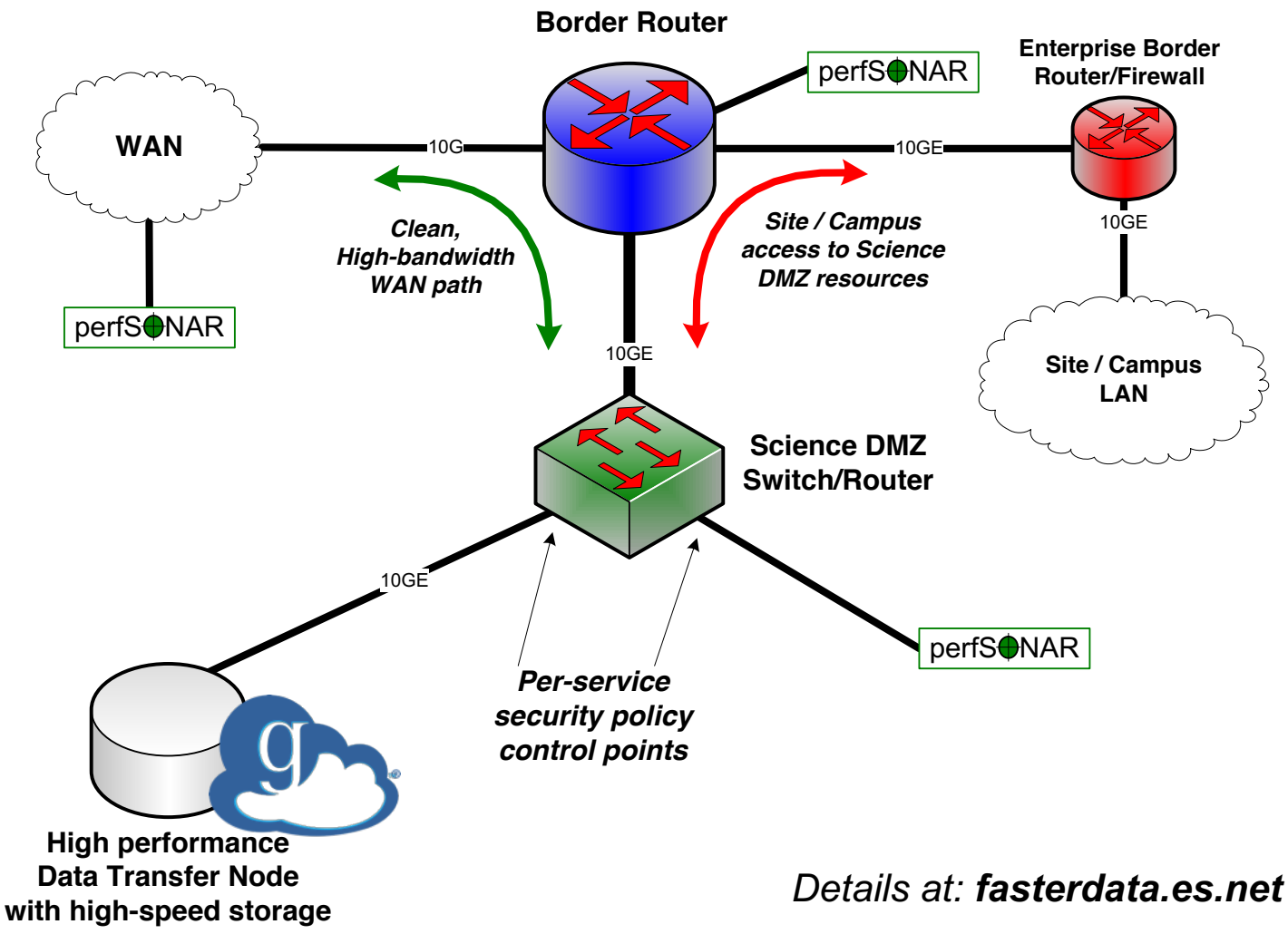
- **Installers do not require admin access**
- **Zero configuration; auto updating**
- **Handles NATs**

Globus Connect Server



- **Enable access for all users with local accounts**
- **Create endpoint on practically any filesystem**
- **Native packages: RPMs and DEBs**

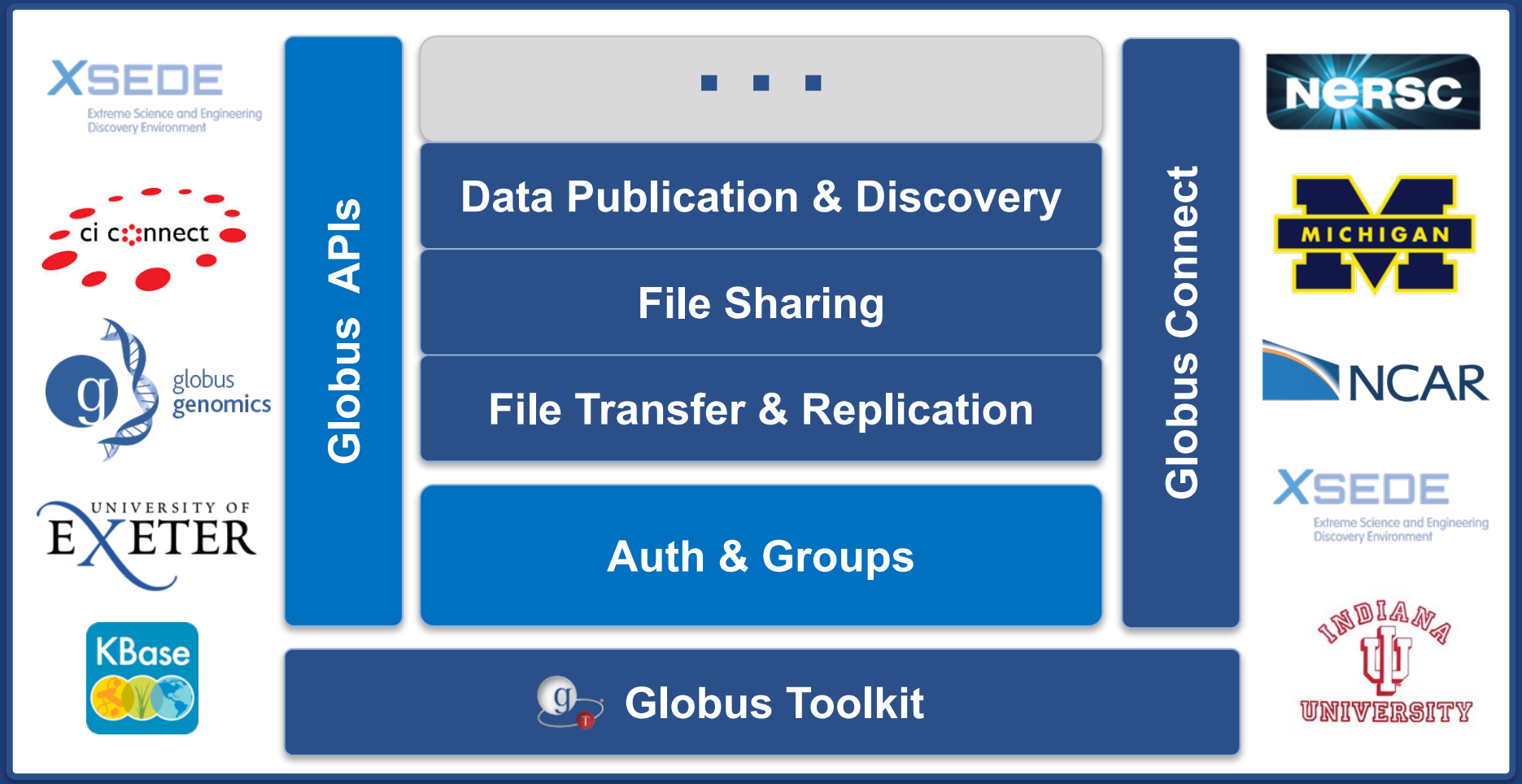
Best-practice deployment



Storage connectors

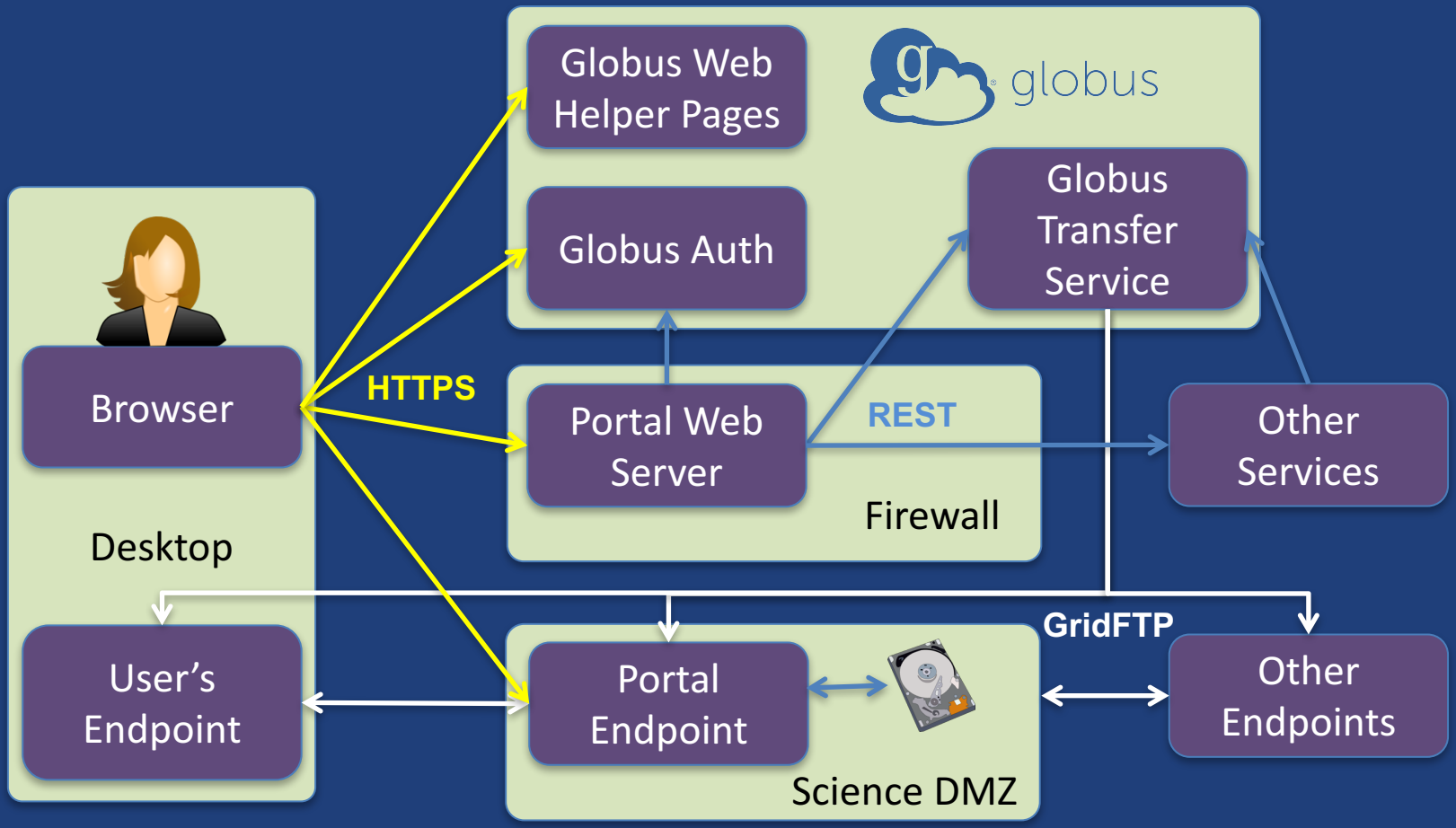
- **Standard storage connectors (Posix)**
  - Linux, Windows, MacOS
  - Lustre, GPFS, OrangeFS, etc.
- **Premium storage connectors**
  - HPSS
  - HDFS
  - S3
  - Ceph RadosGW (S3 API)
  - Spectra Logic BlackPearl
  - Google Drive (coming soon)

Globus Platform-as-a-Service



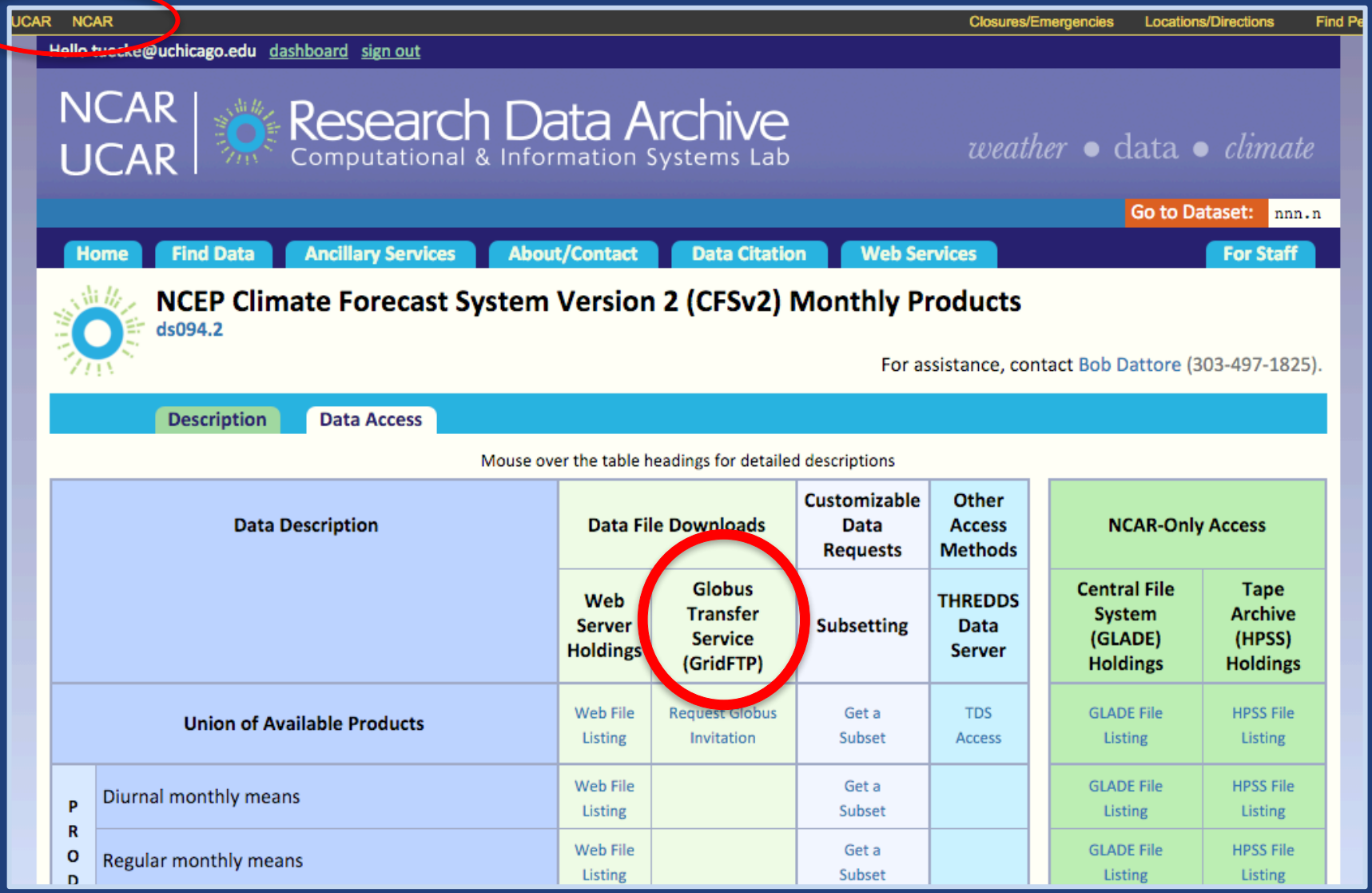
API (and other) documentation: **docs.globus.org**

Building research apps with the Globus Platform-as-a-Service

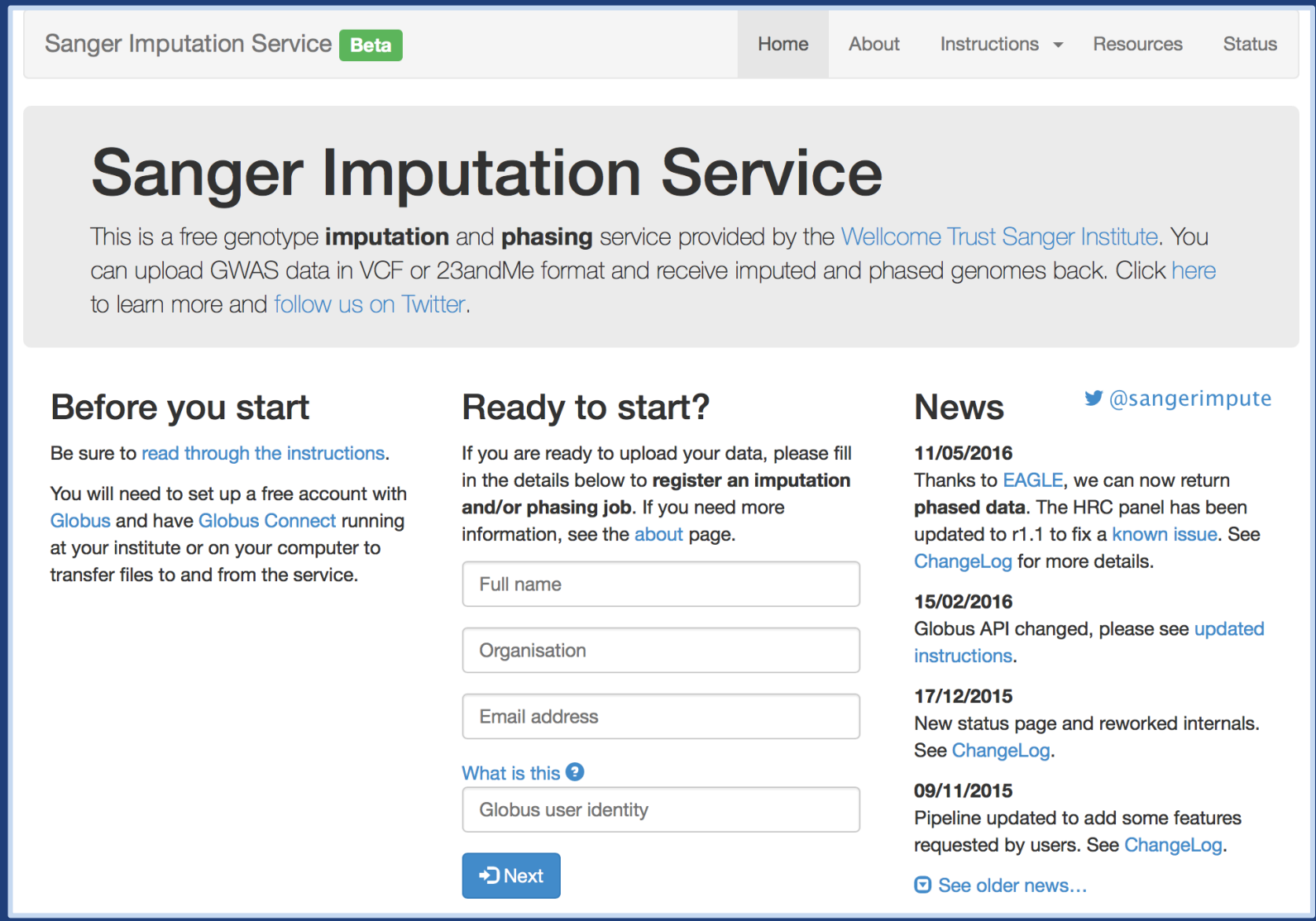


- **Move portal storage into Science DMZ, with Globus endpoint**
- **Leave Portal Web server behind firewall**
- **Globus handles the security and data heavy lifting**

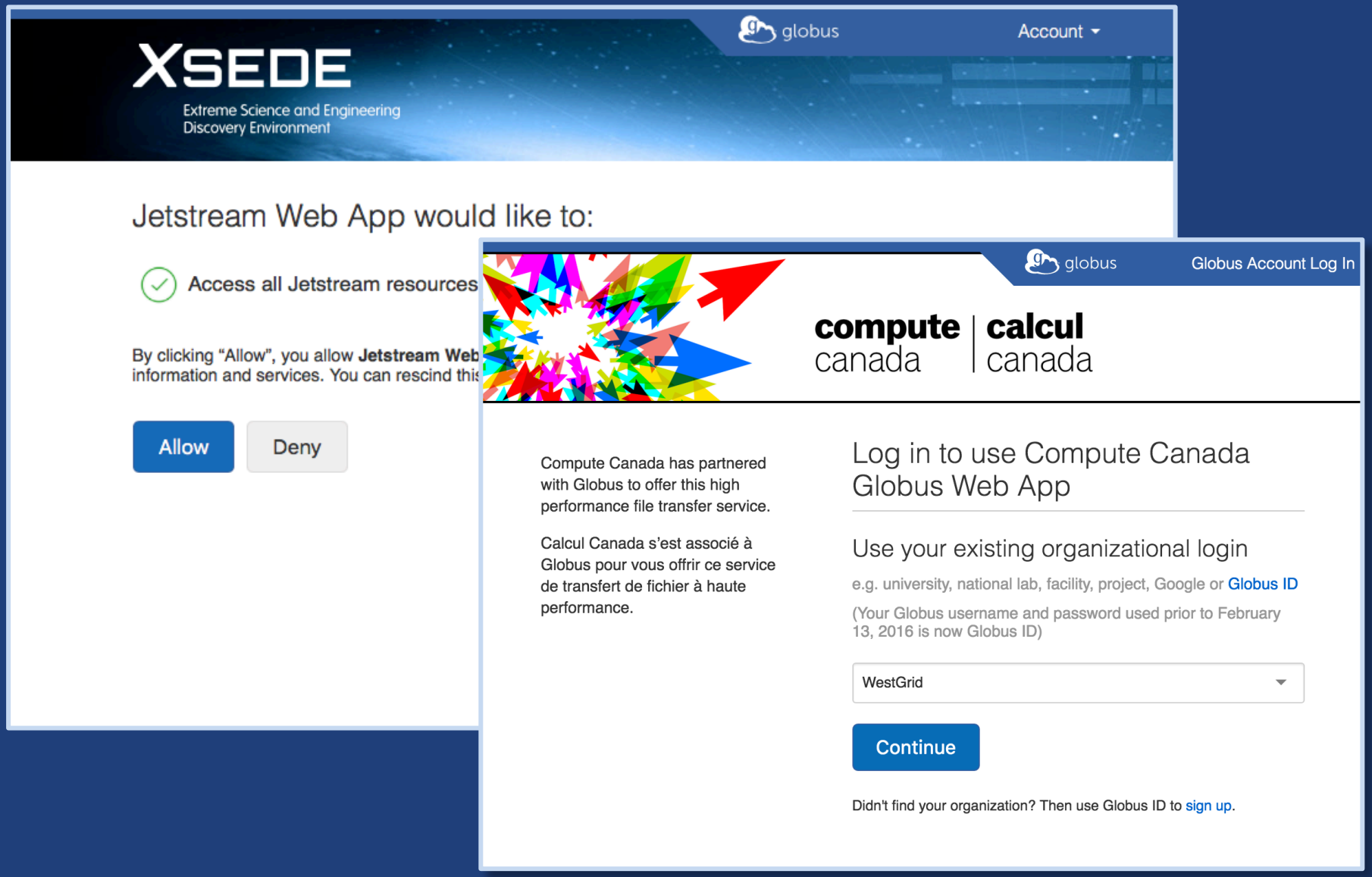
Data App: NCAR RDA



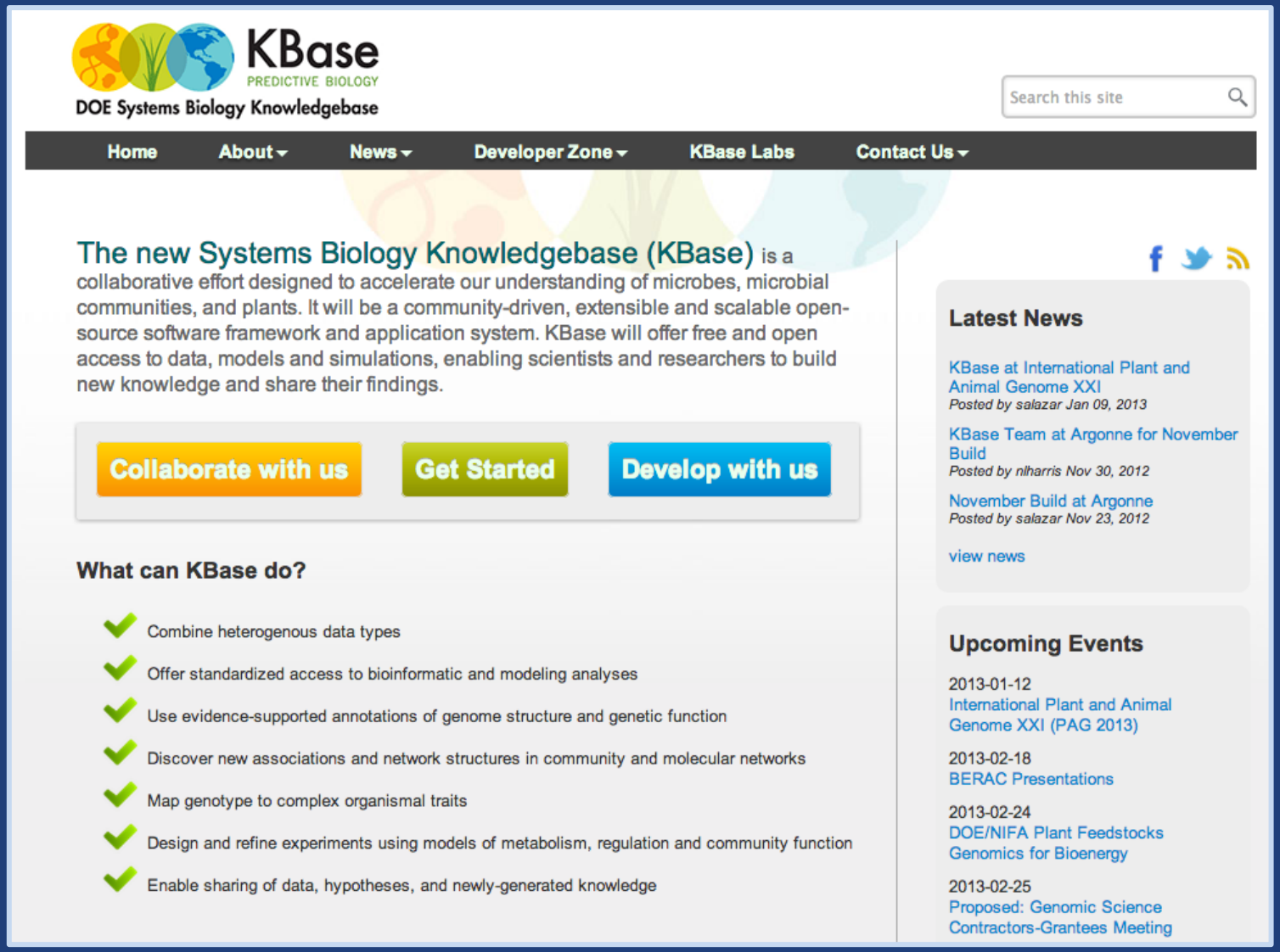
Analysis App: Wellcome Sanger



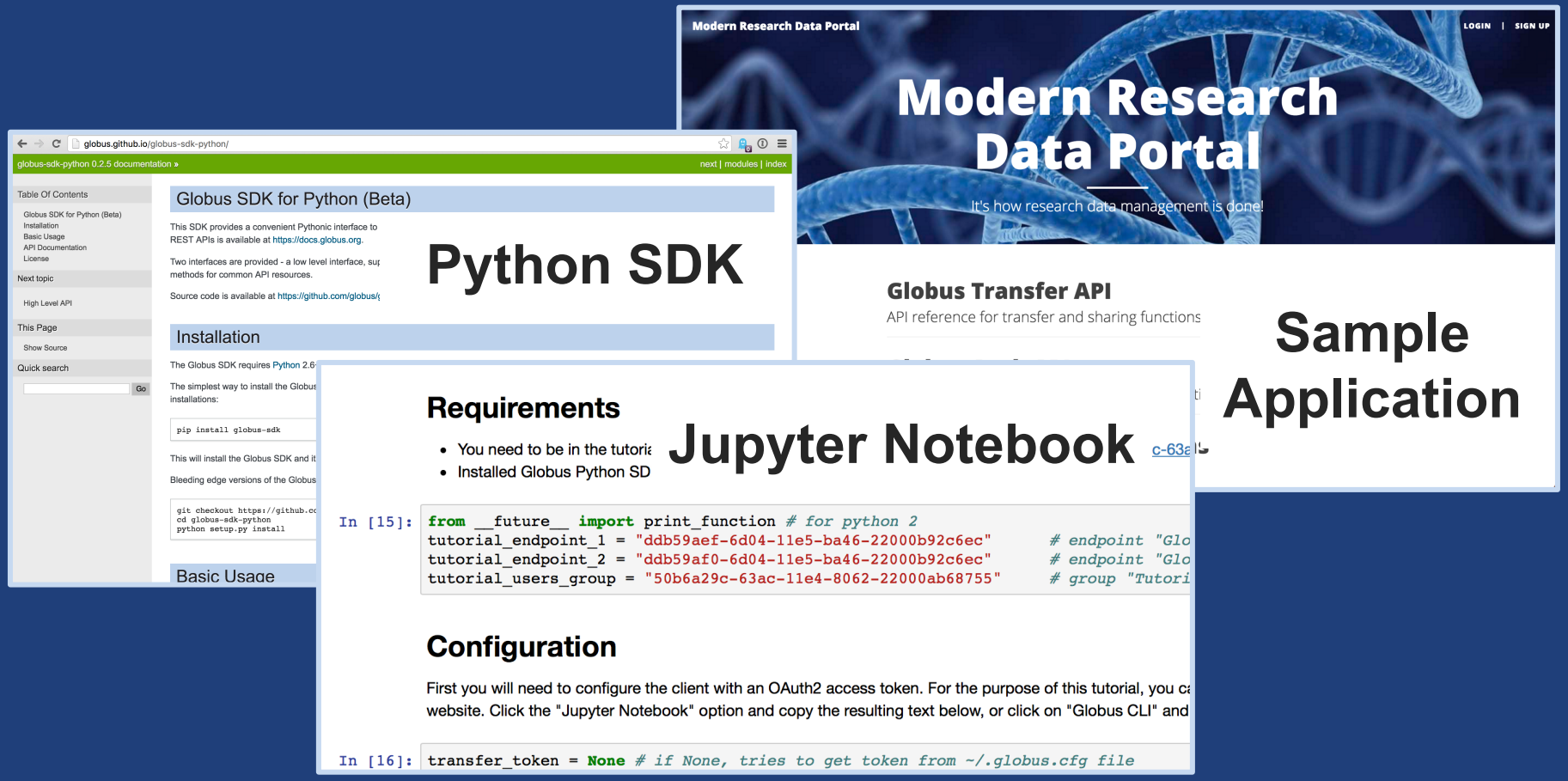
Globus PaaS: National Resource Access



Globus PaaS: Identity Management



Globus PaaS developer resources



docs.globus.org/api

github.com/globus

Demonstration
- Command Line Interface
- Globus PaaS

Globus Milestones

November 18, 2010
Globus Online launches

270 users

April 26, 2011
Globus Connect solves the "last mile" problem

1,000 users

June 4, 2011

January 23, 2012
NCSA selects Globus Online for Blue Waters

June 20, 2012
Globus Online wins R&D 100 Award

January 22, 2013
Globus Online accepted as XSEDE production service

10,000 users

June 13, 2013

November 18, 2013
Globus file sharing announced at SC13

February 2014
NERSC and NCAR become first paying Globus subscribers

March 12, 2015
Management console provides admins with dynamic real-time view into transfer activity

100PB moved

June 26, 2015

February 11, 2016
Globus Auth débuts, providing immediate access to 100,000's of users with existing credentials

October 14, 2016

200PB moved

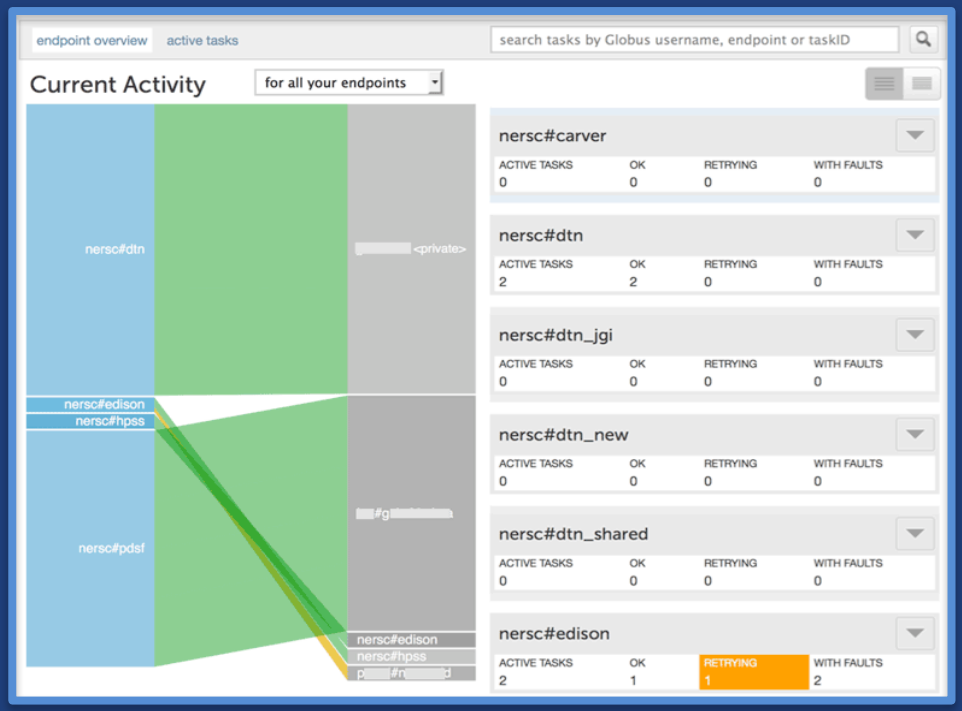
December 3, 2016

50,000 users

Our vision for the future

Globus Subscriptions

- **Standard Subscription**
  - Shared endpoints
  - Data publication
  - HTTPS support*
  - Management console
  - Usage reporting
  - Priority support
  - Application integration
- **Branded Web Site**
- **Premium Storage Connectors**
  - Amazon S3, Ceph, HPSS, Spectra, Google Drive*, …
- **Alternate Identity Provider (InCommon is standard)**



* When available

Demonstration
- Management Console

Thank you to our sponsors

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

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

Argonne
NATIONAL LABORATORY

Thank you to our users...

5 major services	**245 PB** transferred	**40 Bn** files processed	**53,000** 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	**130** federated campus identities

...and thank YOU, our subscribers!

NERSC

NEW YORK UNIVERSITY

Berkeley
UNIVERSITY OF CALIFORNIA

JOHNS HOPKINS
UNIVERSITY

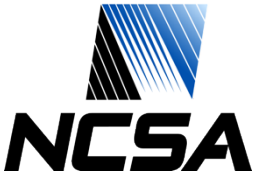
CORNELL
UNIVERSITY

UF | UNIVERSITY of FLORIDA

MICHIGAN STATE
UNIVERSITY

Yale

THE UNIVERSITY OF
CHICAGO



Engage with us

Join the Globus community

- Access the service: **globus.org/login**
- Create endpoints: **globus.org/app/endpoints**
- Learn more: **docs.globus.org**
- Engage: **www.globus.org/mailing-lists**
- Need help? **support@globus.org**
- Subscribe: **globus.org/subscriptions**
- Follow us: **@globusonline**

Thank You.

vas@uchicago.edu