globus online

# Research Data Management

www.globusonline.org

Rachana Ananthakrishnan

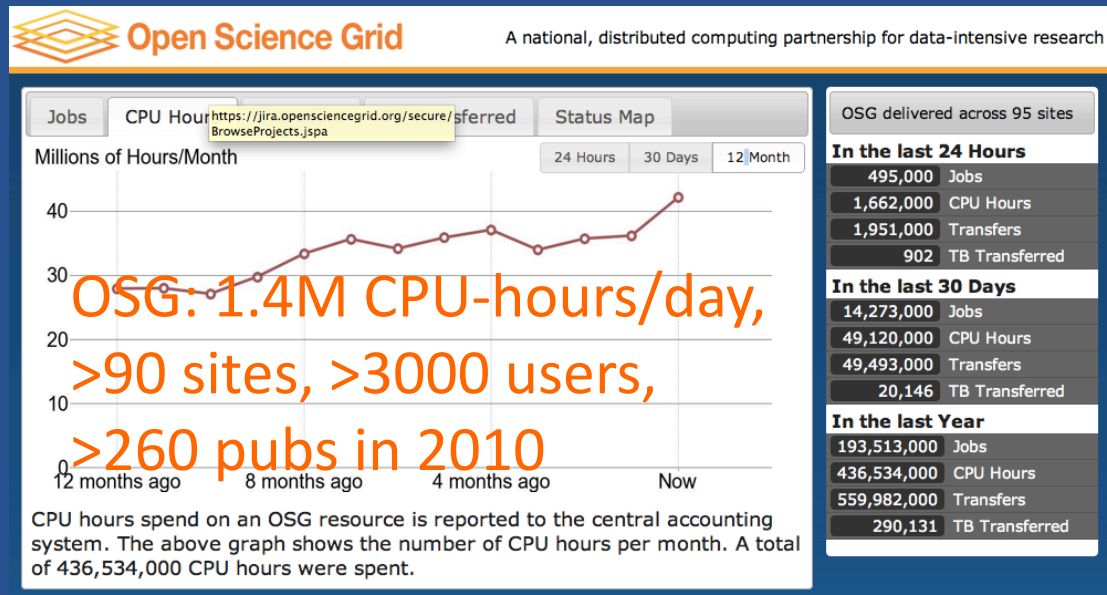
University of Chicago & Argonne National Lab

# We started with technology proven in many large-scale grids

GridFTP
GRAM
MyProxy
GSI-OpenSSH
…

# Big science has achieved big successes with advanced community services


LIGO: 1 PB data in last science run, distributed worldwide




OSG: 1.4M CPU-hours/day, >90 sites, >3000 users, >260 pubs in 2010

- Substantial teams
- Sustained effort
- Leverage common technology
- Application-specific solutions
- Production focus


ESG: 1.2 PB climate data delivered to 23,000 users; 600+ pubs

Community services built on Globus Toolkit software

# But small and medium science is suffering

- Data deluge
- Ad-hoc solutions
- Inadequate software, hardware & IT staff

# Medium science: Dark Energy Survey

- Every night, they receive 100,000 files in Illinois
- They transmit files to Texas for analysis …
  then move results back to Illinois …
  and make them available to users
- Process must be reliable, routine, and efficient
- The cyberinfrastructure team is not large!

Blanco 4m on Cerro Tololo

Image credit: Roger Smith/NOAO/AURA/NSF

# Time-consuming Tasks in Research

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature
- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
- …

# Excerpts from ESNet reports

- “Transfers often take longer than expected based on available network capacities”
- “Lack of an easy to use interface to some of the high-performance tools”
- “Tools [are] too difficult to install and use”
- “Time and interruption to other work required to supervise large data transfers”
- “Need data transfer tools that are easy to use, well-supported, and permitted by site and facility cybersecurity organizations”

# We envisage a world where data …

… flows **rapidly, reliably, and securely** among:

- experimental facilities,
- online and archival storage,
- computing facilities, and
- remote institutions

# We envisage a world where data …

… is easily integrated into **dynamic datasets** that also include metadata and programs necessary to understand and regenerate it

# We envisage a world where data …

… is readily **discoverable and accessible** to collaborators, regardless of their and the data's location

We believe a new approach is needed to deliver data management infrastructure

**Frictionless**
**Affordable**
**Sustainable**

Like 

... but for science!

# Focusing on “frictionless”, we’ve started to do this with the **Globus Online** service …

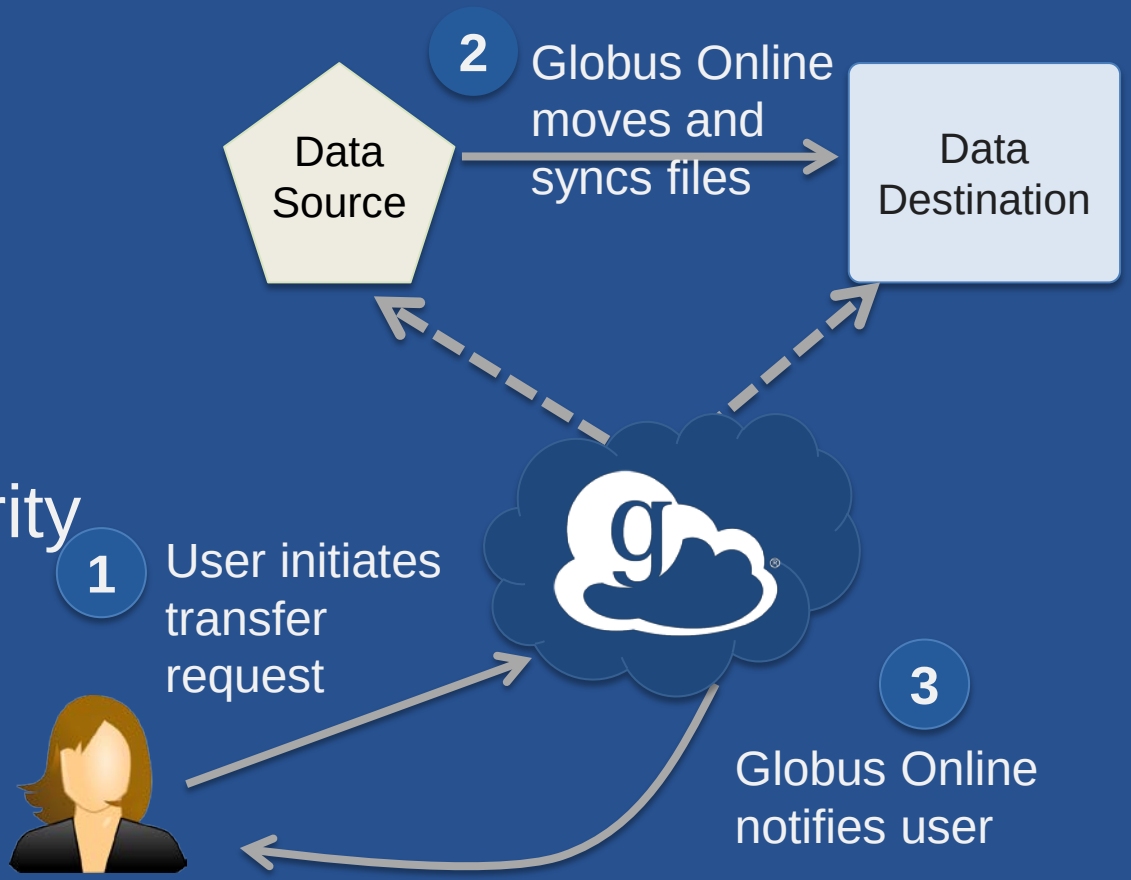
Transfer and sharing of large data sets …

… with dropbox-like characteristics …

… directly from your own storage systems

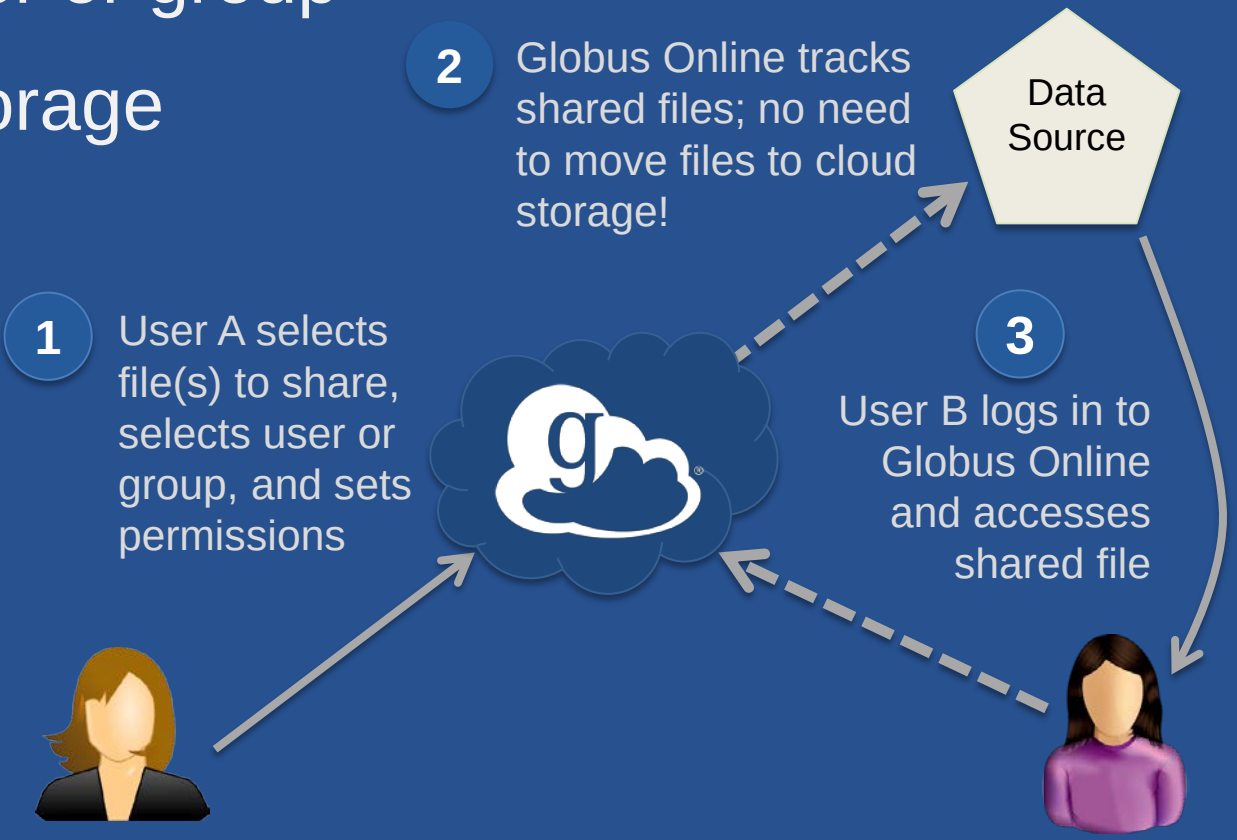
# Reliable, secure, high-performance **file transfer**

- "Fire-and-forget" transfers
- Automatic fault recovery
- Auto tuning
- Seamless security integration



# Simple, secure sharing off existing storage systems

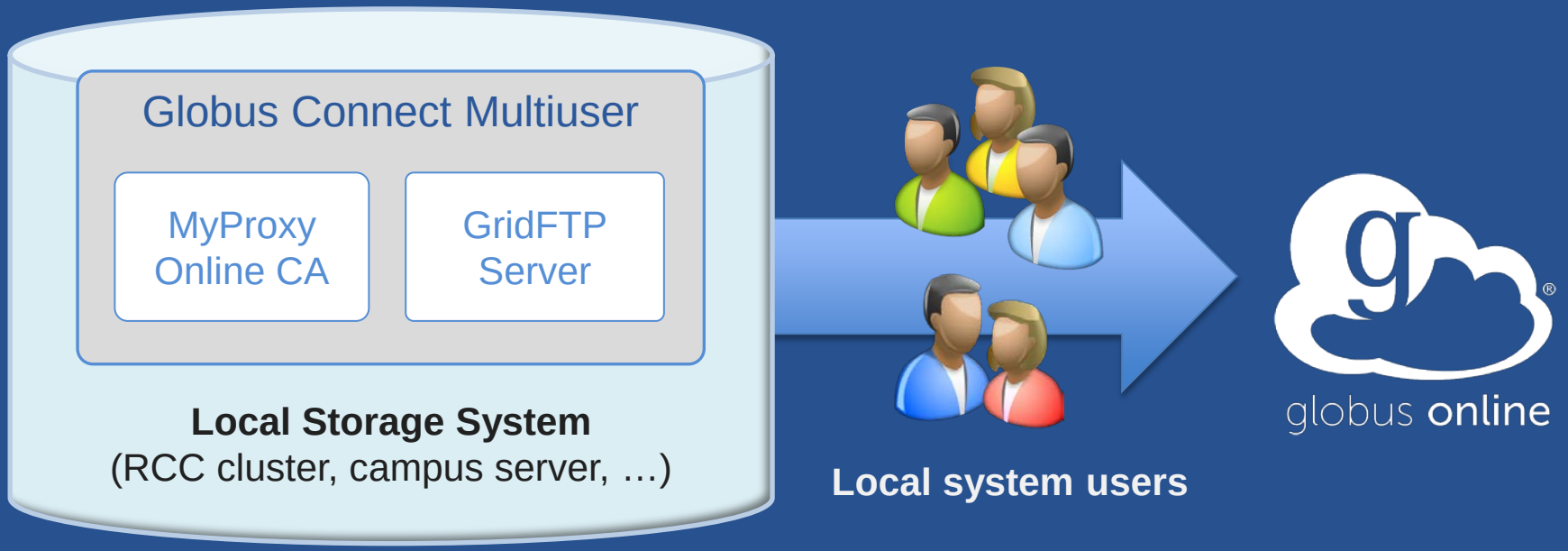
- Easily share large data with any user or group
- No cloud storage required



# Globus Online is SaaS

- Web, command line, and REST interfaces
- Reduced IT operational costs
- New features automatically available
- Consolidated support & troubleshooting
- Easy to add your laptop, server, cluster, supercomputer, etc. with Globus Connect

# Globus Connect Multiuser




- Create endpoint in minutes; no complex GridFTP install
- Enable all users with local accounts to transfer files
- Native packages: RPMs and DEBs
- Also available as part of the Globus Toolkit

# Early adoption is encouraging

globus online

**November 18, 2010**
Globus Online launches

**270** users

**April 12, 2011**
Early users present at GlobusWORLD

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

**June 4, 2011**

**1,000** users


**October 26, 2011**
"MB Transferred" counter débuts, hypnotizing many a researcher

**January 23, 2012**
NCSA selects Globus Online for Blue Waters

**5,000** users

**May 17, 2012**

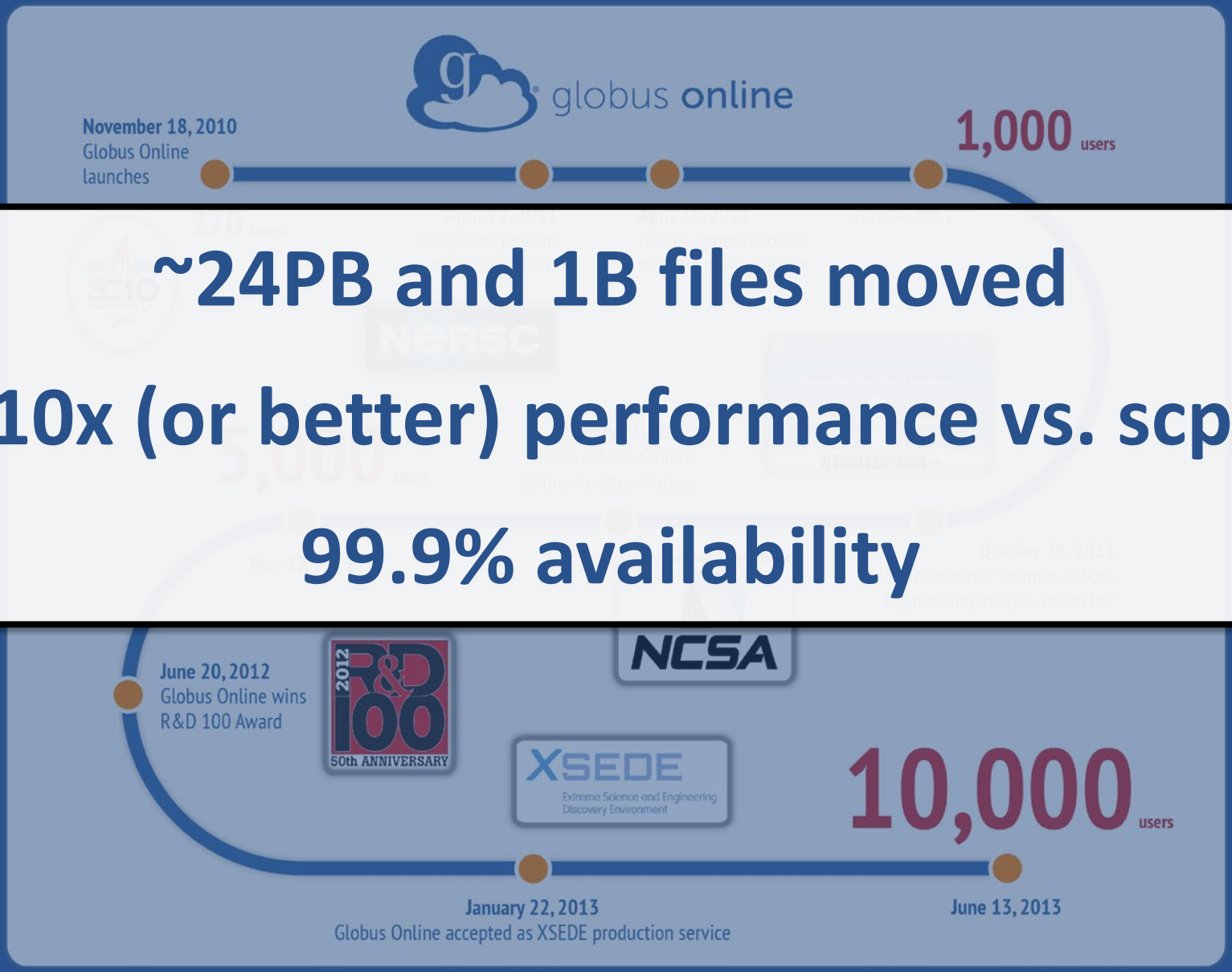
**June 20, 2012**
Globus Online wins R&D 100 Award

**January 22, 2013**
Globus Online accepted as XSEDE production service

**10,000** users

**June 13, 2013**

# Early adoption is encouraging




**~24PB and 1B files moved**

**10x (or better) performance vs. scp**

**99.9% availability**

FOR USERS

- Live Status
- My NERSC
- Getting Started
- Computational Systems
- Data & File Systems
  - Data Management Policies
  - NERSC File Systems
  - HPSS Data Archive
  - Optimizing I/O performance on the Lustre file system
  - I/O Formats
  - Sharing Data
  - Transferring Data
    - Globus Online
    - SCP/SFTP
    - bbcp
    - NERSC FTP Upload Service
  - Unix Groups at NERSC
  - Unix File Permissions
- Network Connections
- Queues and Scheduling
- Job Logs & Analytics
- Training & Tutorials
- Software

Home » For Users » Data & File Systems » Transferring Data » Globus Online

# GLOBUS ONLINE

## Overview

Globus Online addresses the challenges faced by researchers in moving, sharing, and archiving large volumes of data among distributed sites. With Globus Online, you hand-off data movement tasks to a hosted service that manages the entire operation, monitoring performance and errors, retrying failed transfers, correcting problems automatically whenever possible, and reporting status to keep you informed while you focus on your research. Command line and web-based interfaces are available. The command line interface, which requires only ssh to be installed on the client, is the method of choice for grid-based workflows.

TABLE OF CONTENTS

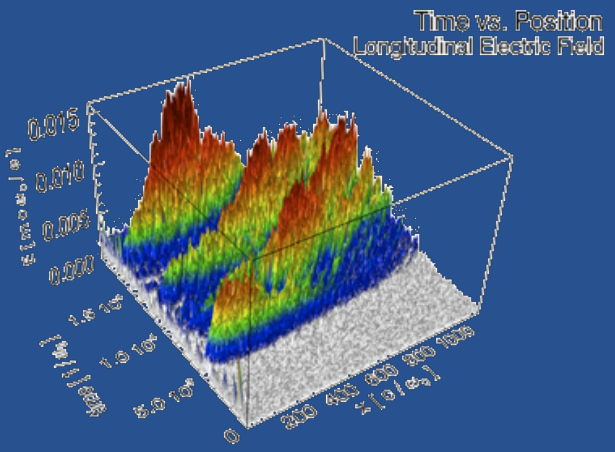


As described below you **register** with Globus Online, and then use the NERSC **endpoint** "nersc#dtn" as well as other sources or destinations. The NERSC endpoints listed are NERSC's **data transfer nodes**, which are tuned especially for WAN data movement tasks. You can activate the NERSC endpoints on Globus Online by simply using your NERSC username and password.

Back to Top

## Availability

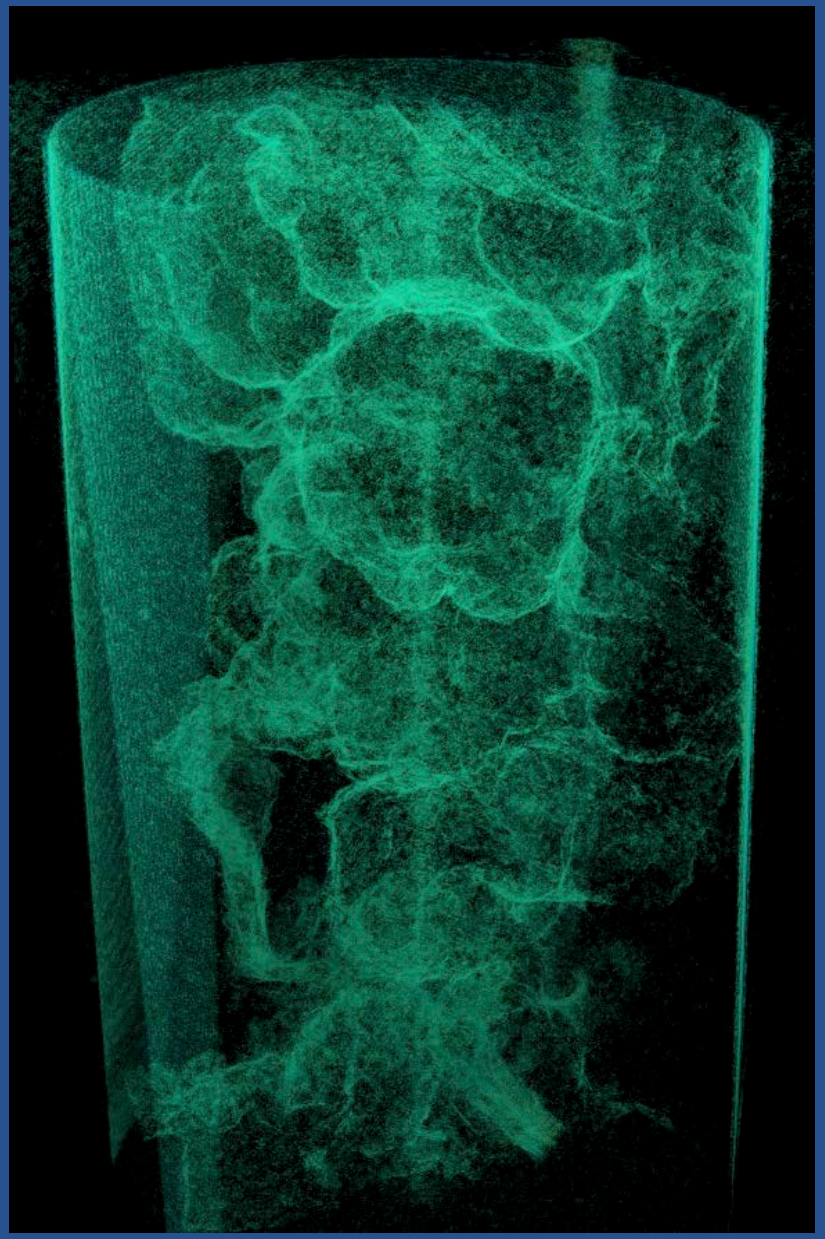
Globus Online is **available** as a free service that any user can sign up for. NERSC's data transfer nodes (DTNs), Hopper, and PDSF are available today as endpoints on Globus Online (in addition to dozens of other endpoints from other sites). If you would like to see a destination added as an endpoint feel free to contact NERSC and/or the staff at that location to get it added to the growing registry of endpoints. You can also add an endpoint on your laptop or local workstation with **Globus Connect**.




B. Winjum (UCLA) moves 900K-file **plasma physics** datasets UCLA →NERSC

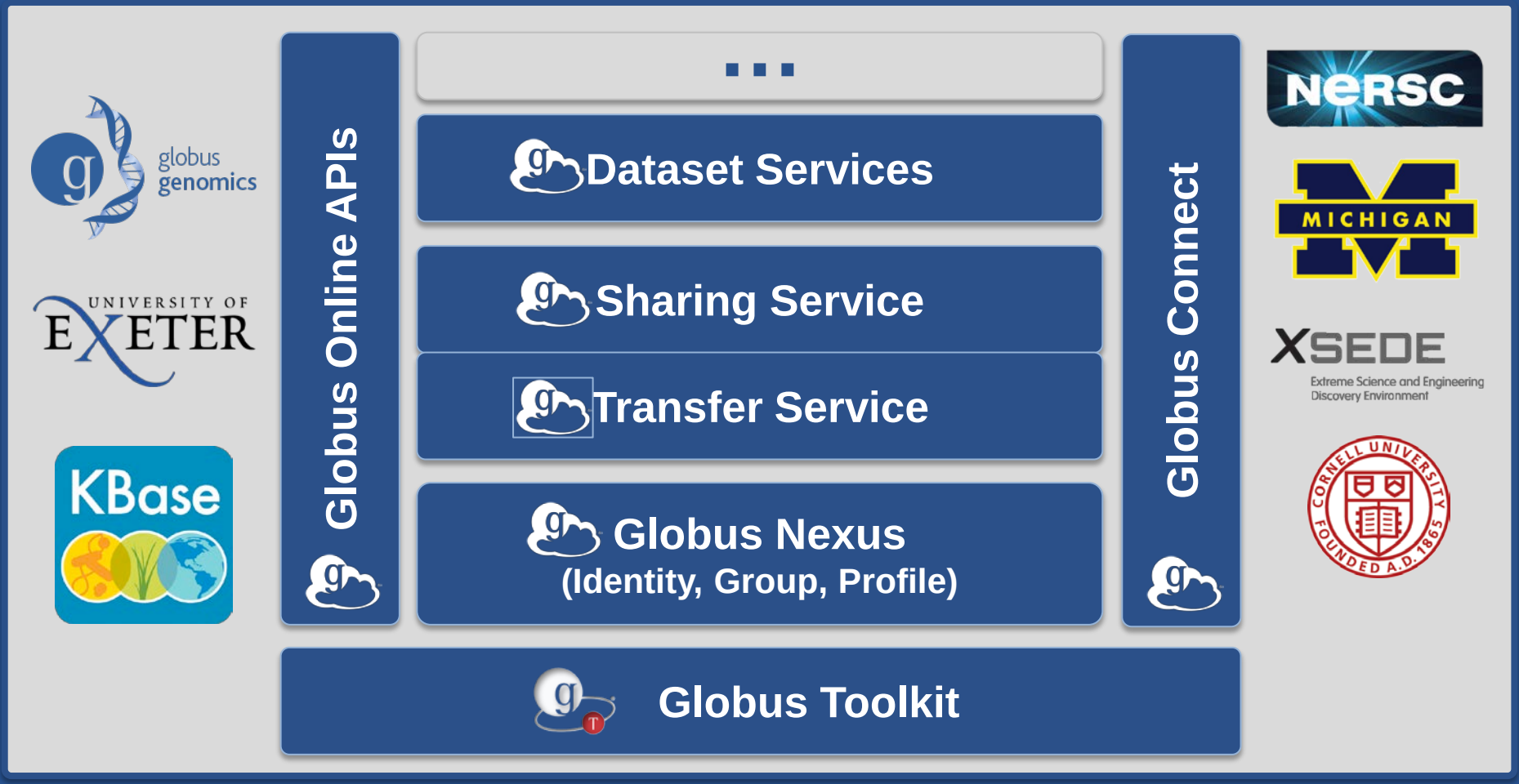
Dan Kozak (Caltech) replicates 1 PB LIGO **astronomy** data for resilience

Erin Miller (PNNL) collects data at Advanced Photon Source, renders at PNNL, and views at ANL



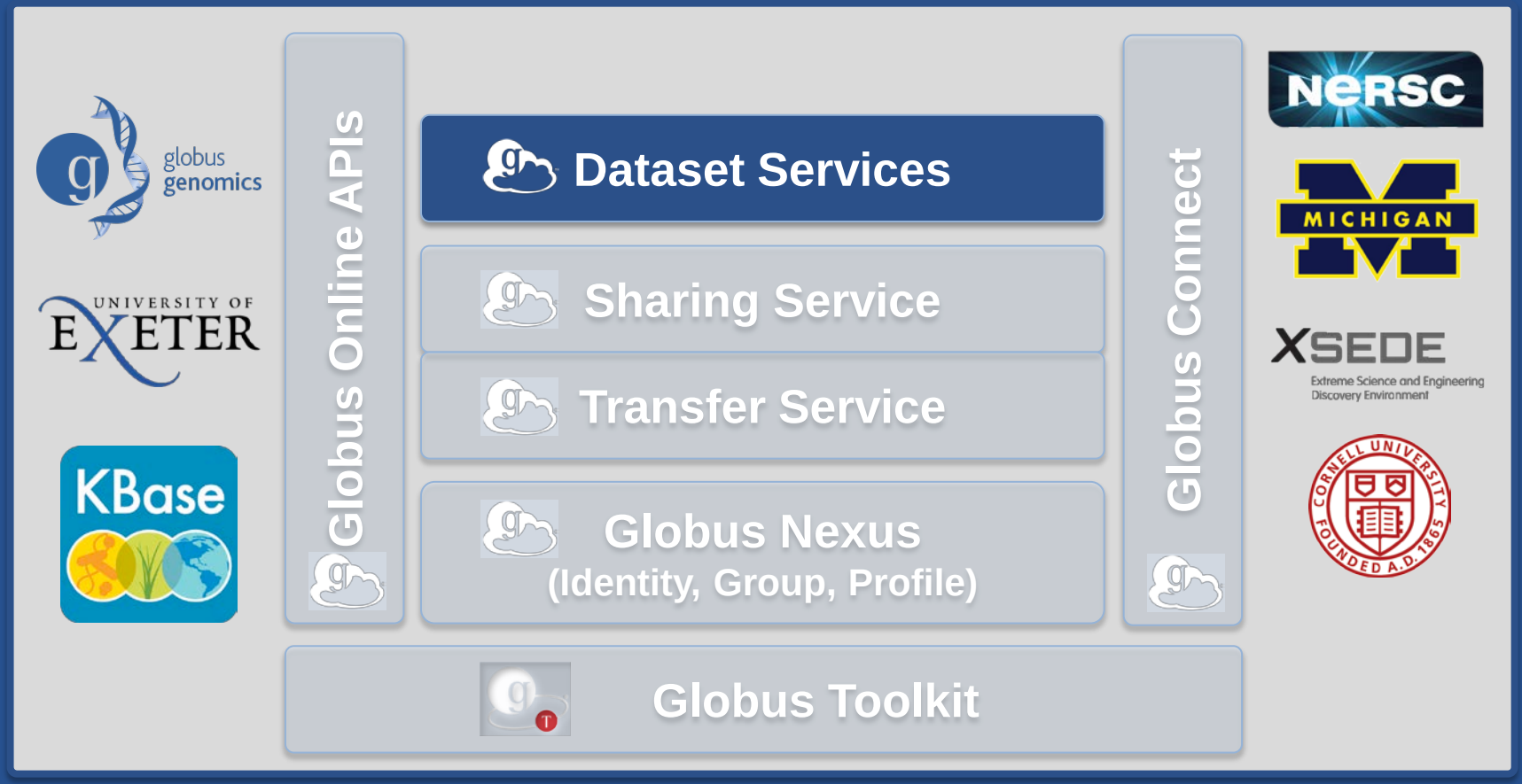
Credit: Kerstin Kleese-van Dam

# Globus Online as a platform



# Early platform adopters

# More capabilities underway …



# Introducing the **dataset**

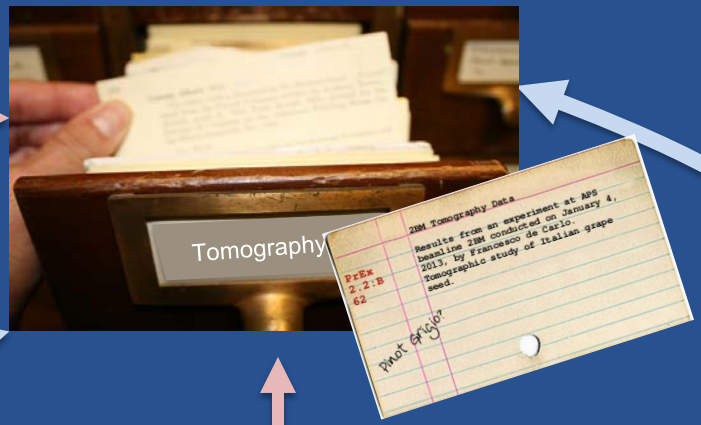
- **Group** data based on use, not location
  - Logical grouping to organize, reorganize, search, and describe usage
- **Tag** with characteristics that reflect content …
  - Capture as much existing information as we can
- …or to reflect current status in investigation
  - Stage of processing, provenance, validation, ..
- **Share** data sets for collaboration
  - Control access to data and metadata
- **Operate** on datasets as units
  - Copy, export, analyze, tag, archive, …

# Expanding Globus Online services

- Ingest and publication
  - Imagine a DropBox that not only replicates, but also extracts metadata, catalogs, converts
- Cataloging
  - Virtual views of data based on user-defined and/or automatically extracted metadata
- Integration with computation
  - Associate computational procedures, orchestrate application, catalog results, record provenance

Annotate, share

browse, search




Catalog derived products

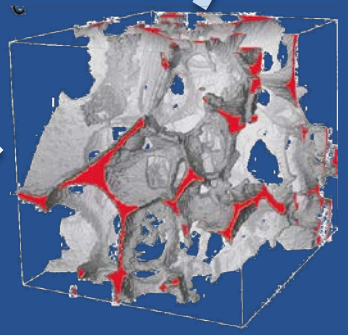
Populate catalog(s)

mydata42

Record provenance

Define dataset
Infer type
Extract metadata

analyze

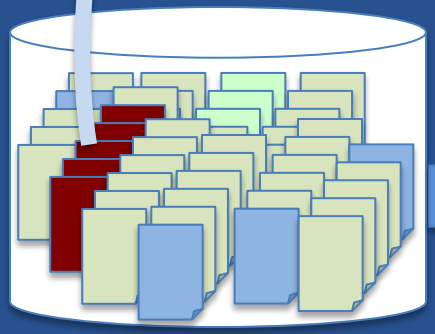


**owner**: Francesco
**type**: 3dtomo
**format**: HDF5
**beamline**: 2BM

## Organization

## Orchestration

Locate datasets
Access files



transfer/schedule

We believe a new approach is needed to deliver data management infrastructure

**Frictionless**

**Affordable**

**Sustainable**

# We've got a handle on "frictionless"

- Web interface, REST API, command line
- InCommon, Oauth, OpenID, X.509, …
- Credential management
- Group definition and management
- Transfer management and optimization
- Reliability via transfer retries
- One-click "Globus Connect" install
- 5-minute Globus Connect Multiuser install

# “Affordable” and “sustainable”?

Common expectation is **either**:

- High-priced commercial software (with generally higher levels of quality)

**Or**:

- Free, open source software (with generally lower levels of quality)

We aim to offer the best of all worlds!

We are a non-profit service provider to the non-profit research community

We are a non-profit service provider to the non-profit research community

Our challenge:

Sustainability

# Globus Online Provider Plans

Support ongoing operations

Offer value-added capabilities

Engage more closely with users

# Provider Plans offer…

- Endpoint management console
- Usage reporting
- MSS optimizations
- Globus Plus subscriptions
- Branded web sites
- Alternate identity provider

Starting at $10k/year

# Researchers may use Globus file transfer for free

- File transfer and synchronization to/from servers
- Personal endpoints with Globus Connect
- Access to shared endpoints created by others

- Globus Plus: $7/month (or $70/year)
  - Create and manage shared endpoints
  - Transfer and sharing between Globus Connect Personal endpoints

# We hope you will join us

XSEDE
Extreme Science and Engineering Discovery Environment

NERSC

HARVARD

MICHIGAN

UNIVERSITY OF EXETER

UNIVERSITY of WASHINGTON

NCSA

Carnegie Mellon University

NGS

APS physics

NORTHWESTERN UNIVERSITY 1851

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

ISI
Information Sciences Institute

KSU

NASA

INDIANA UNIVERSITY

Fermilab

EMORY

ESnet
Energy Sciences Network

BERKELEY LAB
Lawrence Berkeley National Laboratory

Los Alamos
NATIONAL LABORATORY
EST. 1943

CERN

CORNELL
UNIVERSITY

Cal

Ole Miss

THE UNIVERSITY OF CHICAGO

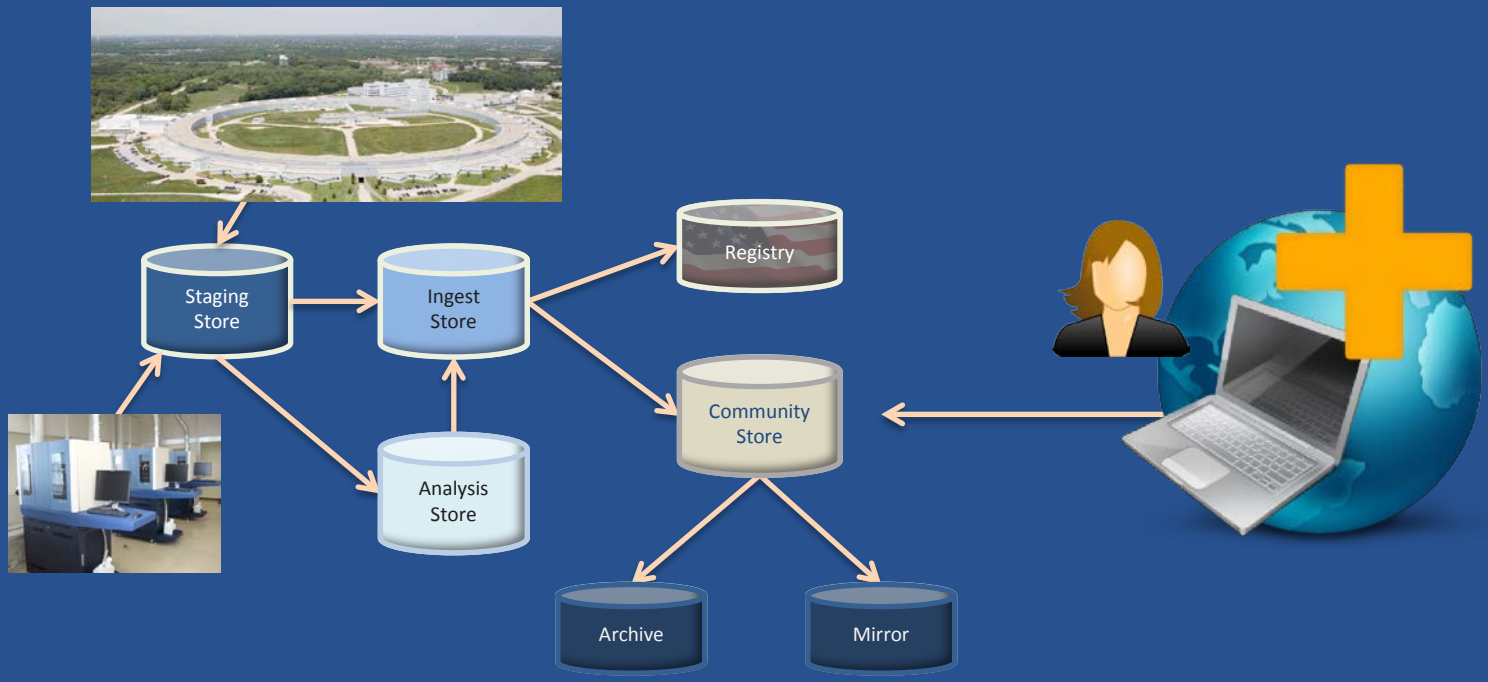
NEW YORK UNIVERSITY

Argonne
NATIONAL LABORATORY

# Provider Plan not required to get started

Use Globus Connect Multiuser to easily connect your resources with Globus Online

Go to: **globusonline.org/gcmu**



# Our research is supported by:


U.S. DEPARTMENT OF
ENERGY




THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY

powered by

# Questions

**Contact**: support@globusonline.org

**Providers**: globusonline.org/provider-plans

**Researchers**: globusonline.org/plus

www.globusonline.org