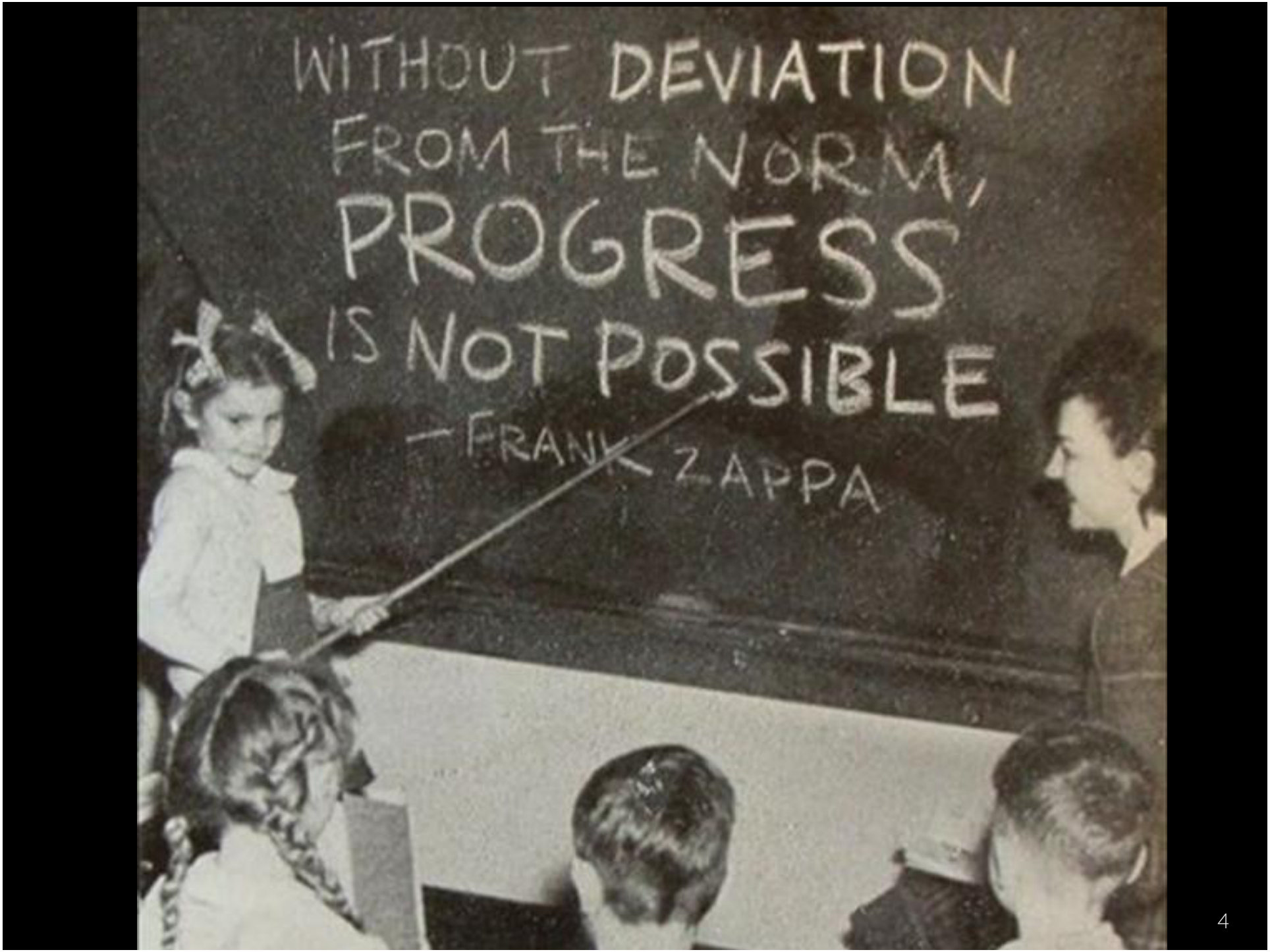
Managing data
throughout the research lifecycle
using Globus

GlobusWorld 2015
Keynote

#globus15

@globusonline


data deluge


WITHOUT DEVIATION
FROM THE NORM,
PROGRESS
IS NOT POSSIBLE
— FRANK ZAPPA

Globus ...

… simplifies
research data management

… using cloud-hosted services

Why cloud-hosted services matter

Think: amazon

- Logic runs in the cloud
- Interactions occur via Web or API
- Actions occur in the world

Simplicity, reliability, economies of scale

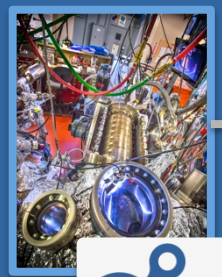
Globus delivers …

… data transfer, sharing, publication, and discovery

… on storage chosen by you

Managing the research data lifecycle

Light Source



Globus transfers files reliably, securely

2

Compute Facility

Transfer

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

1 PI initiates transfer request; or requested automatically by script or science gateway

3 Share

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

6 Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)

Publication Repository

- SaaS → Only a web browser required
- Access using your campus credentials
- Globus monitors and informs throughout

5 Researcher logs in and accesses shared files; no local account required; download via Globus

Publish

6

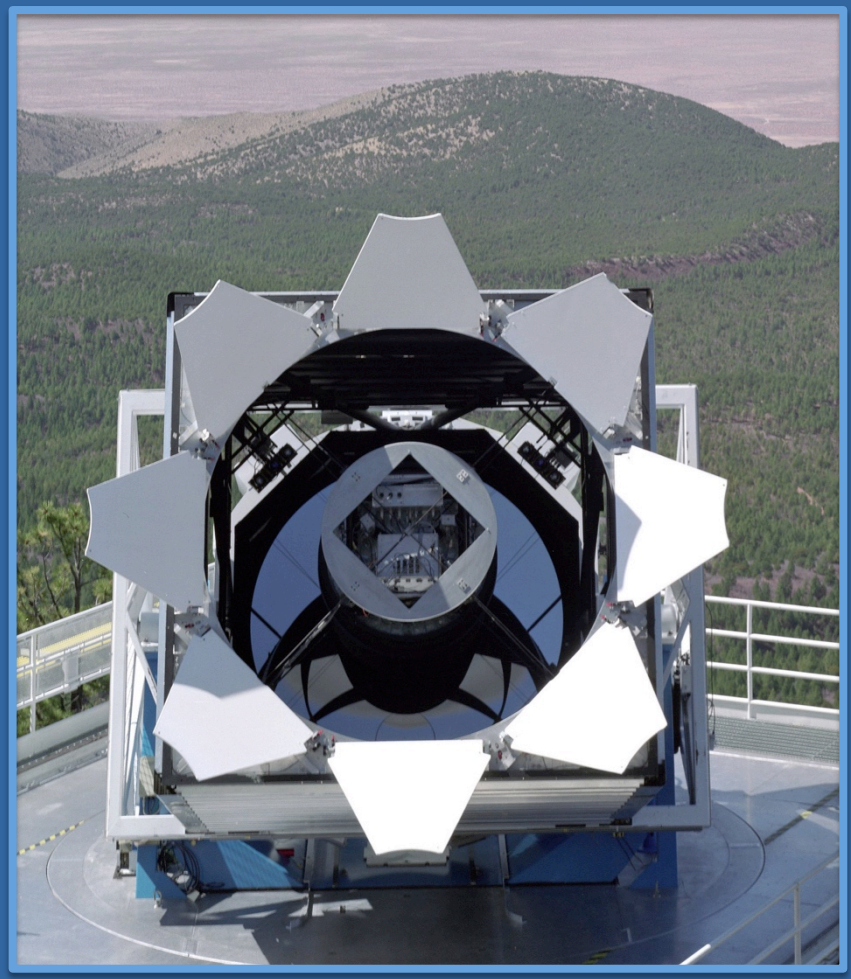
Personal Computer

8

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

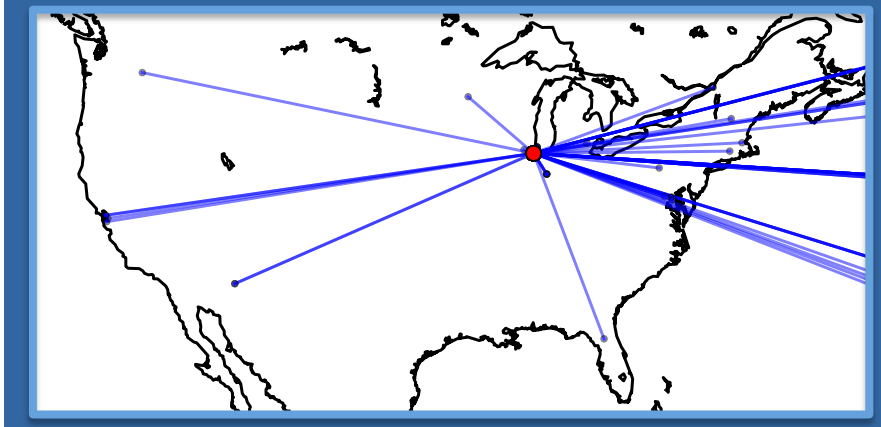
Discover

400,000 transfers during Jan-Feb 2015 from SDSS repository at NYU

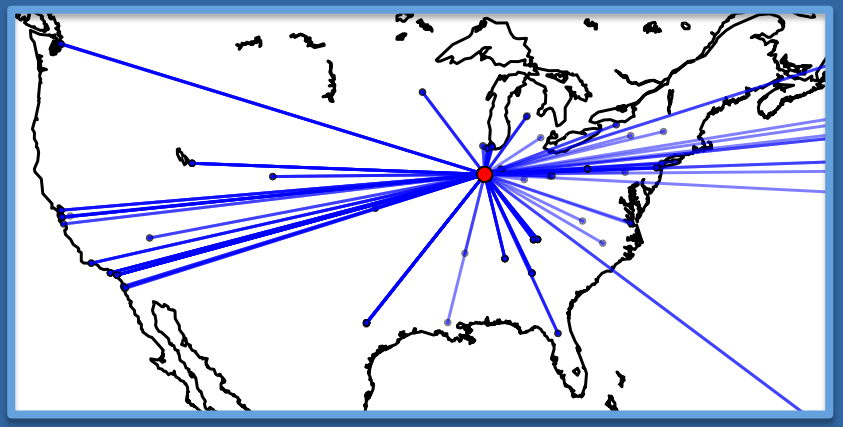


Sloan Digital Sky Survey
Source: University of Utah

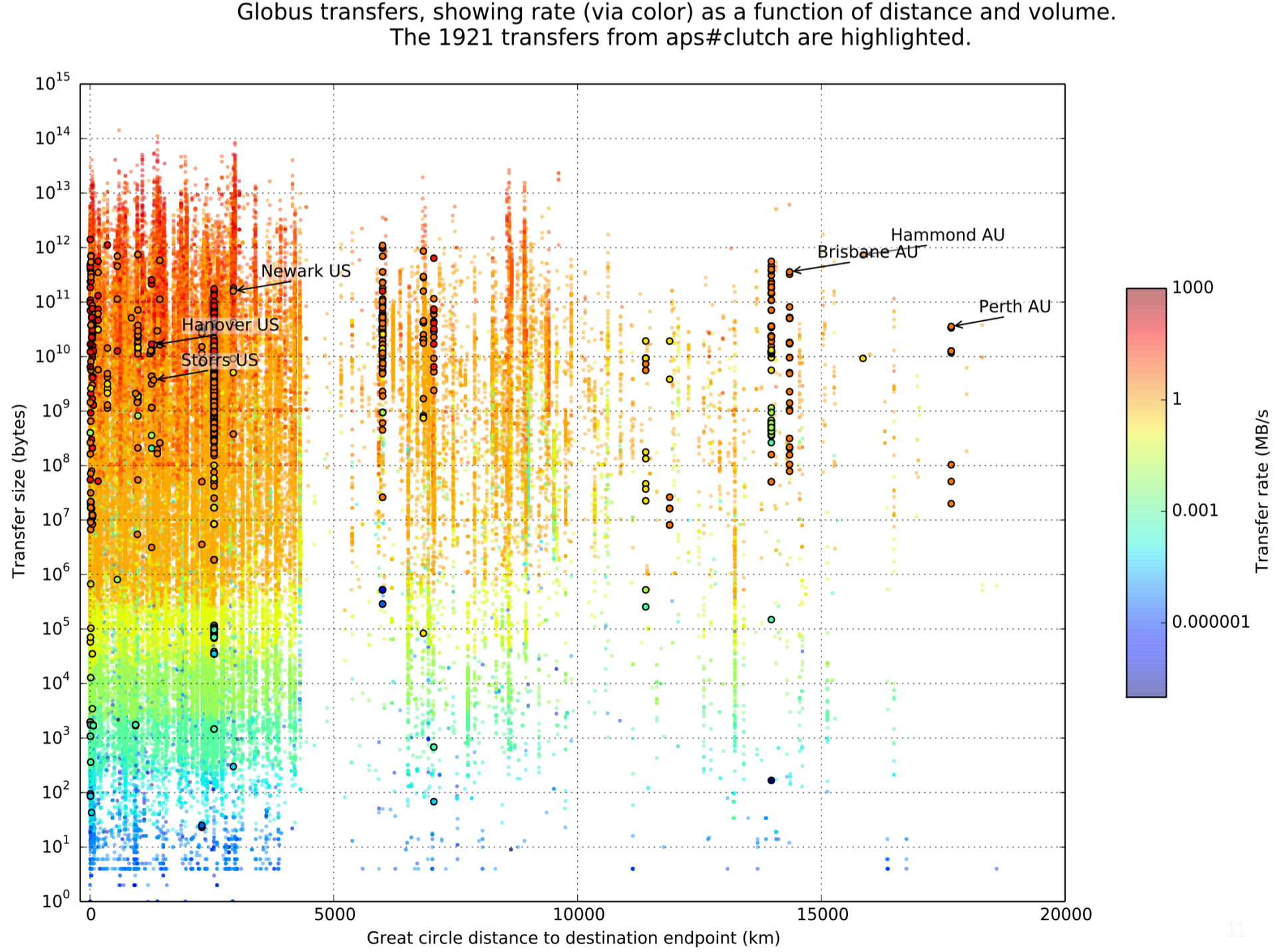
People are moving data worldwide



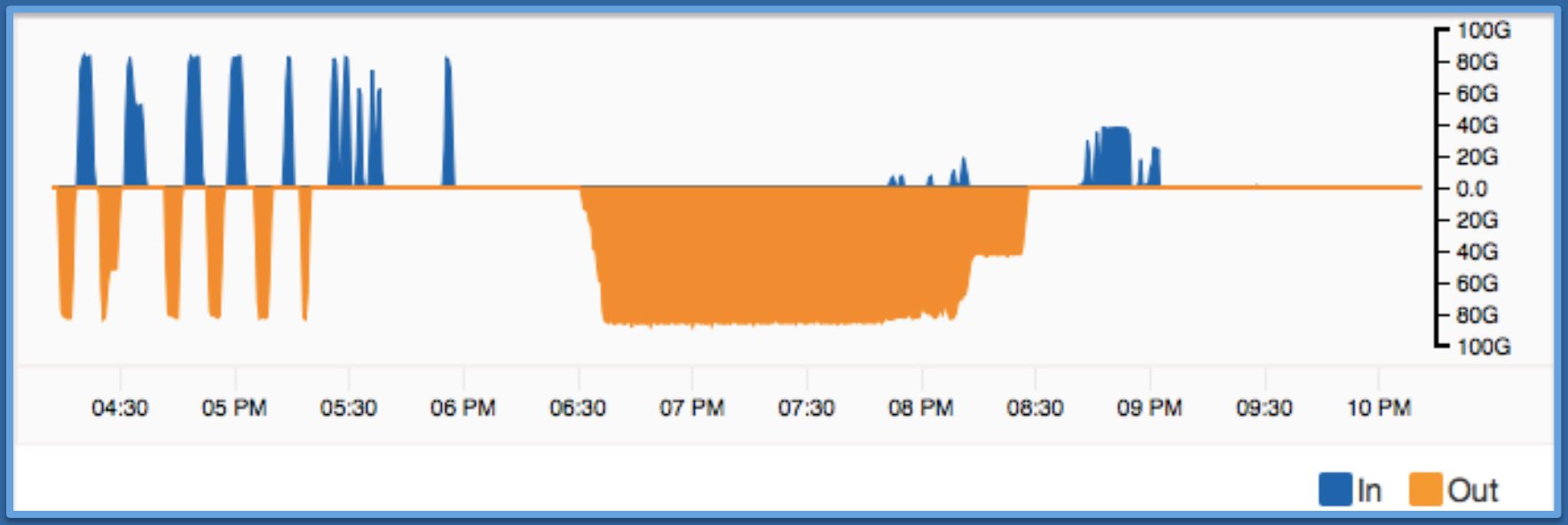
Advanced Photon Source
2,275 transfers to 119 destinations



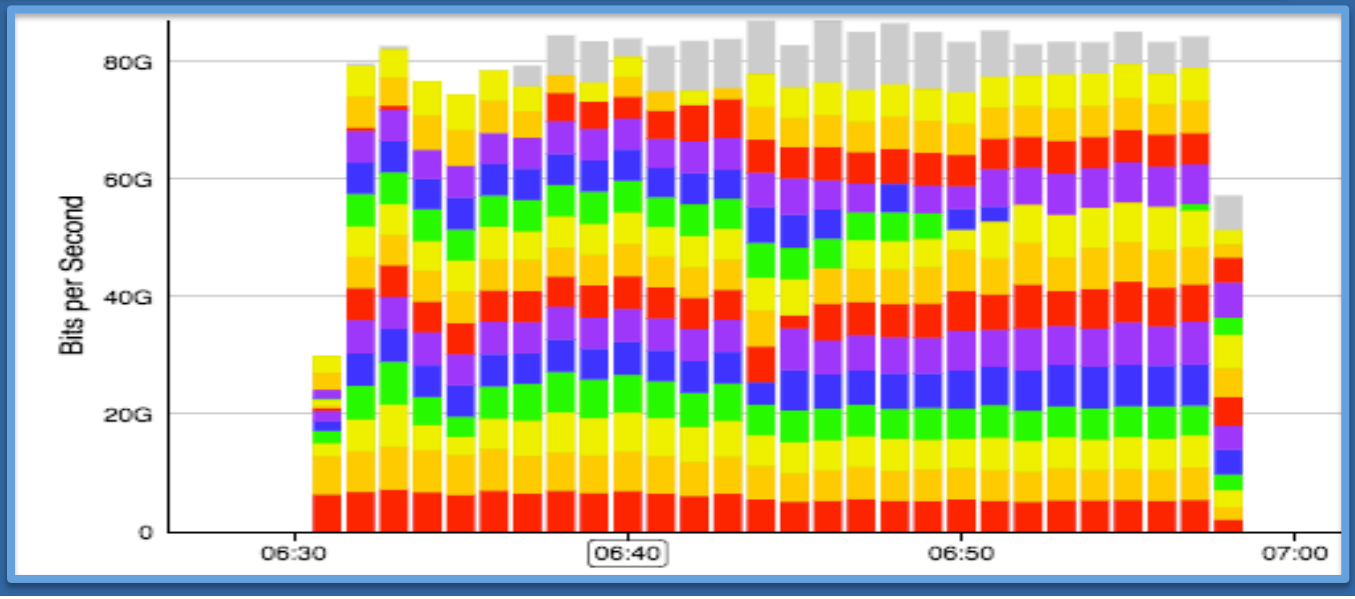
Blue Waters at NCSA
79,525 transfers to 300 destinations


Globus transfers, showing rate (via color) as a function of distance and volume.
The 1921 transfers from aps#clutch are highlighted.
Transfer size (bytes)
Great circle distance to destination endpoint (km)
Transfer rate (MB/s
Newark US
Hanover US
Storrs US
Brisbane AU
Hammond AU
Perth AU
1000
1
0.001
0.000001
0
5000
10000
15000
20000

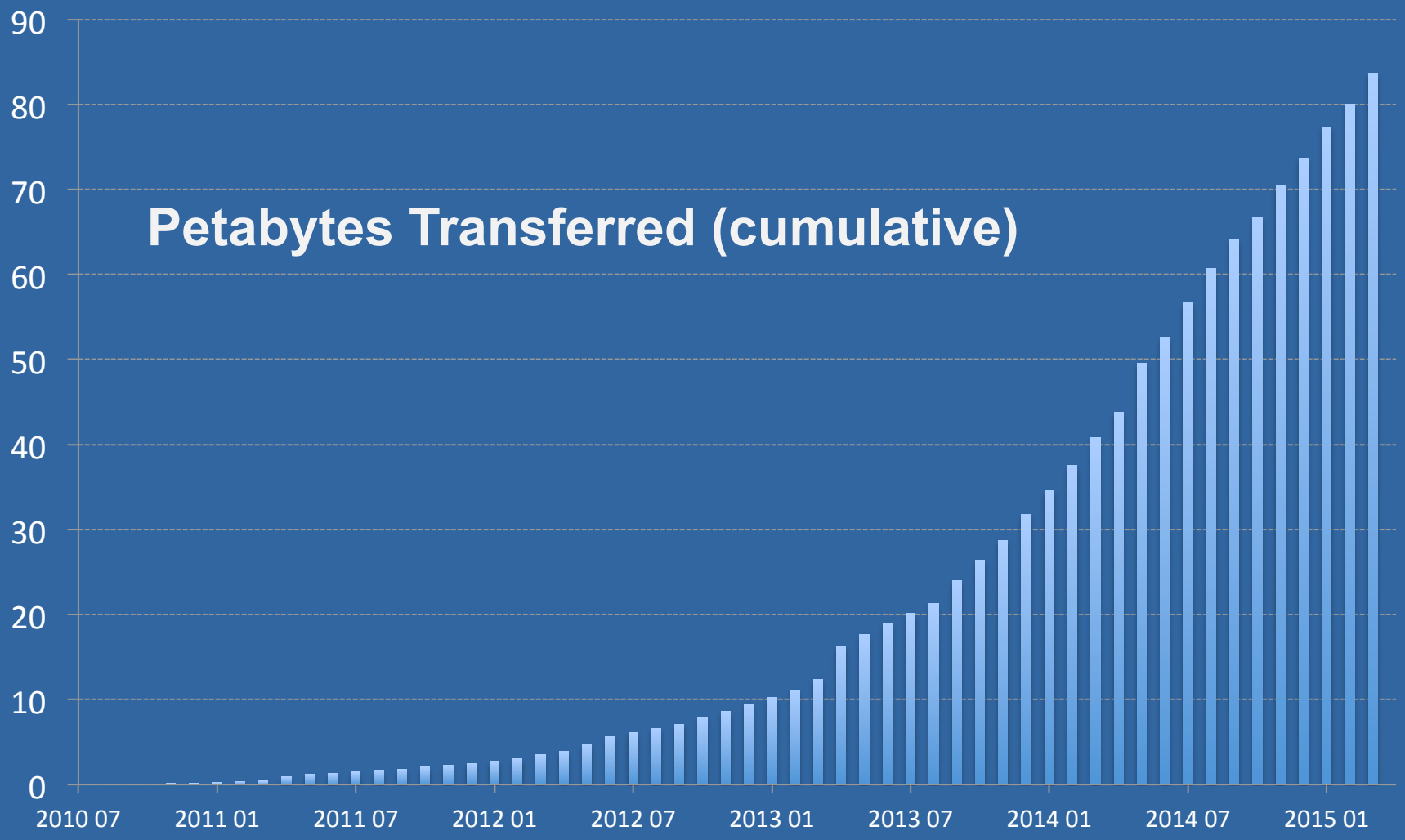
85 Gbps sustained disk-to-disk over 100 Gbps network, Ottawa—New Orleans



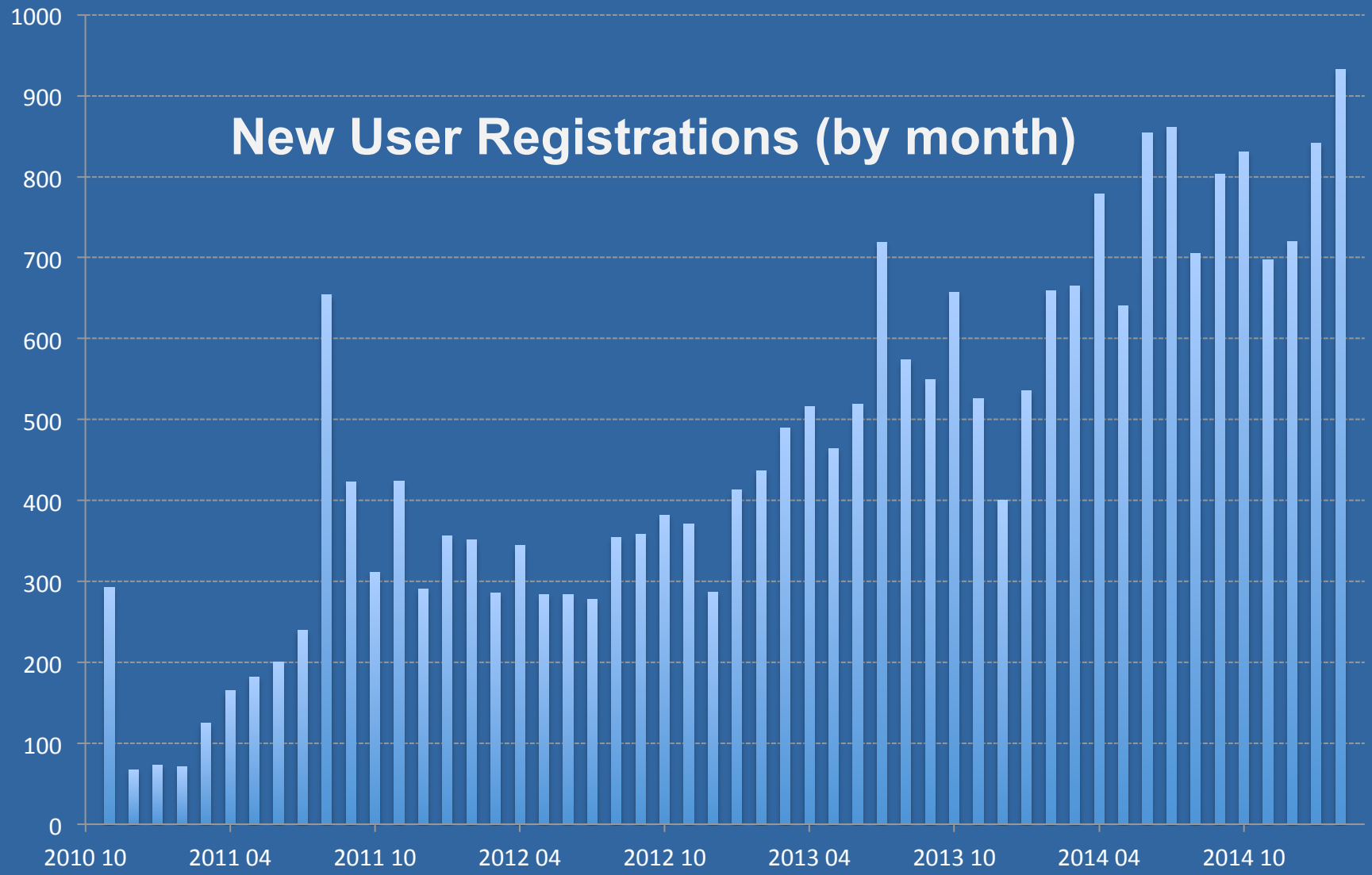
Raj Kettimuthu and team, Argonne Nov 2014



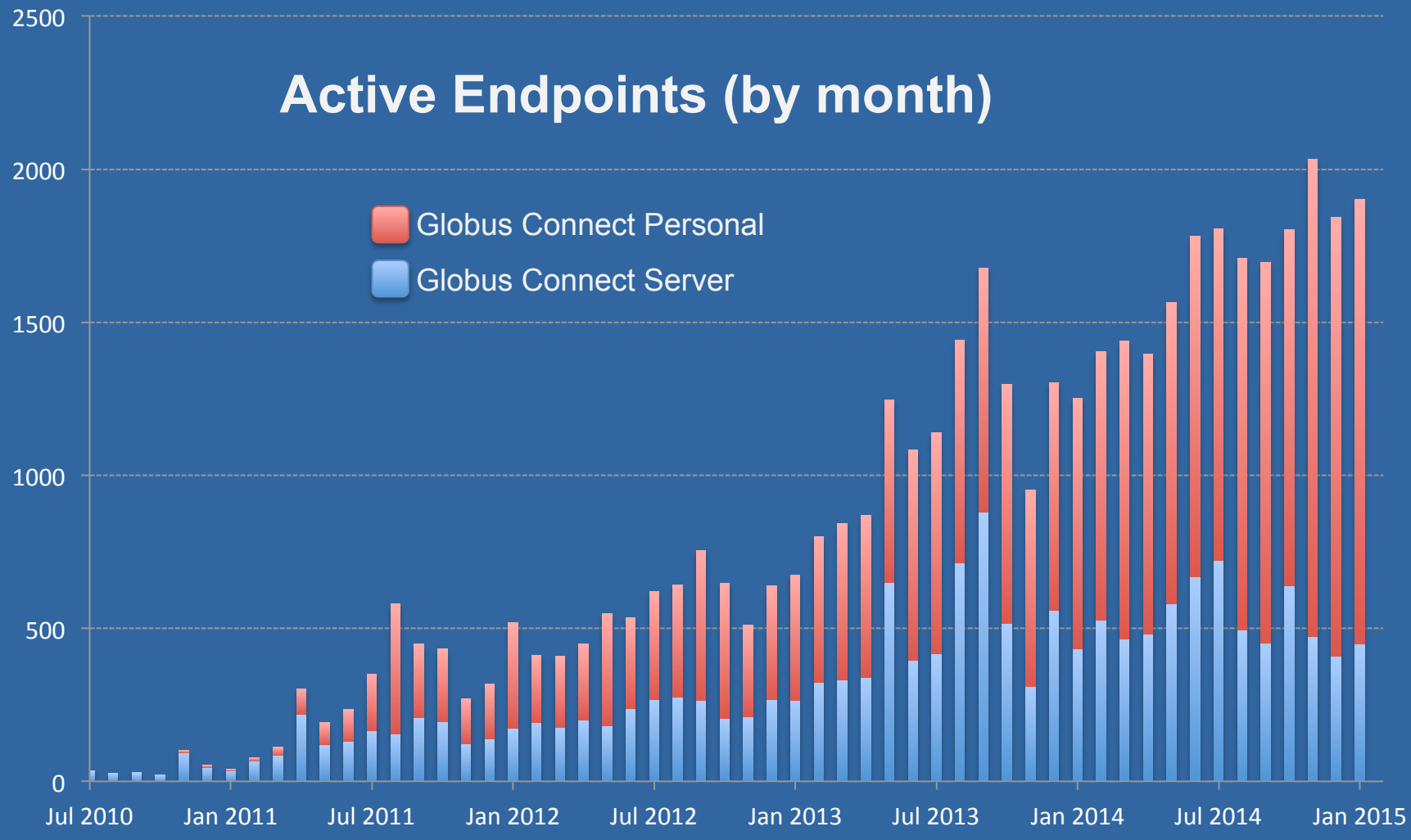
Broad adoption continues



Broad adoption continues



Broad adoption continues



Data publication update

Light Source

Compute Facility

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

2 Globus transfers files reliably, securely

3 PI 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 Researcher 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

Publication Repository

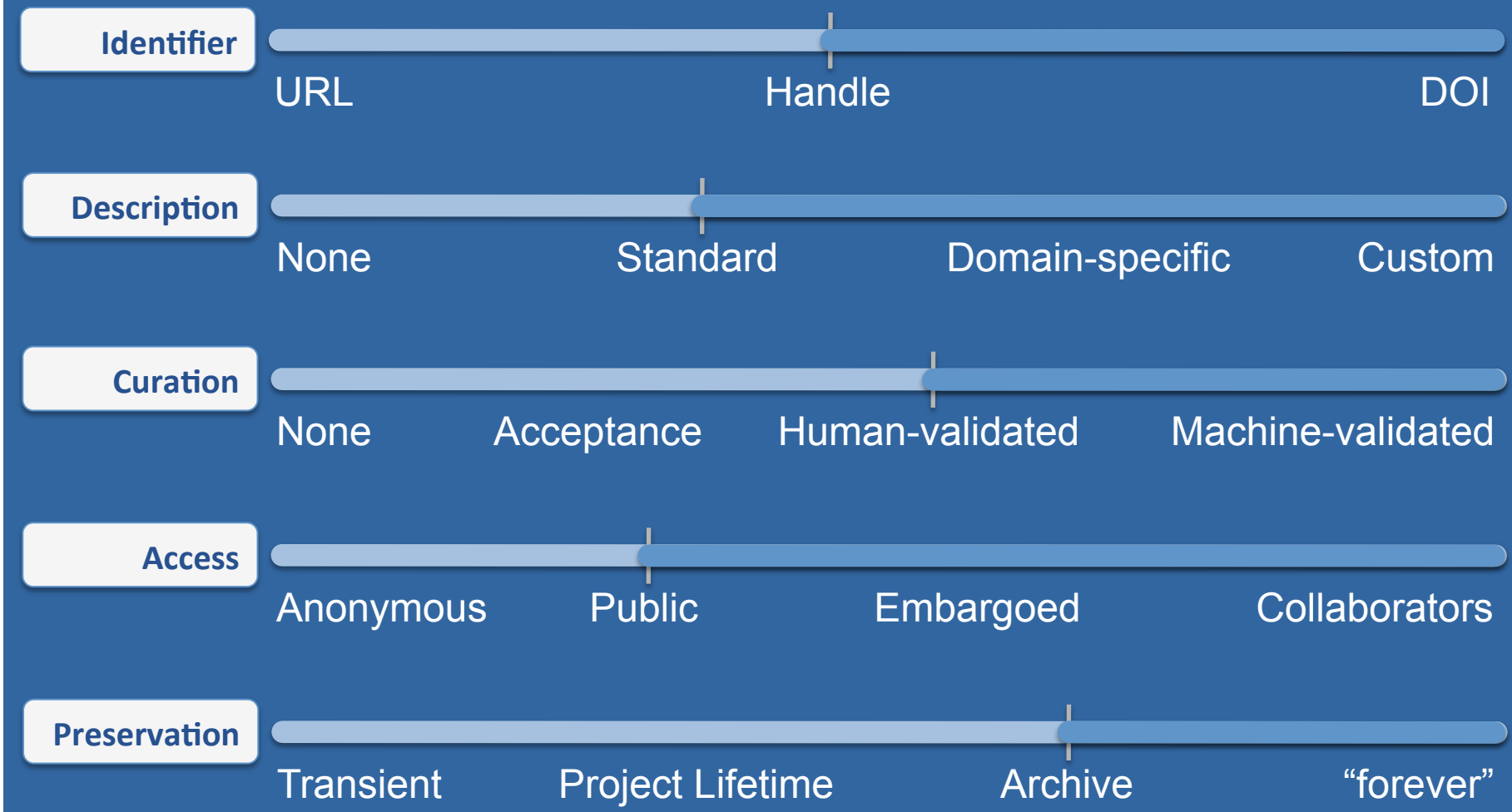
Personal Computer

Publish

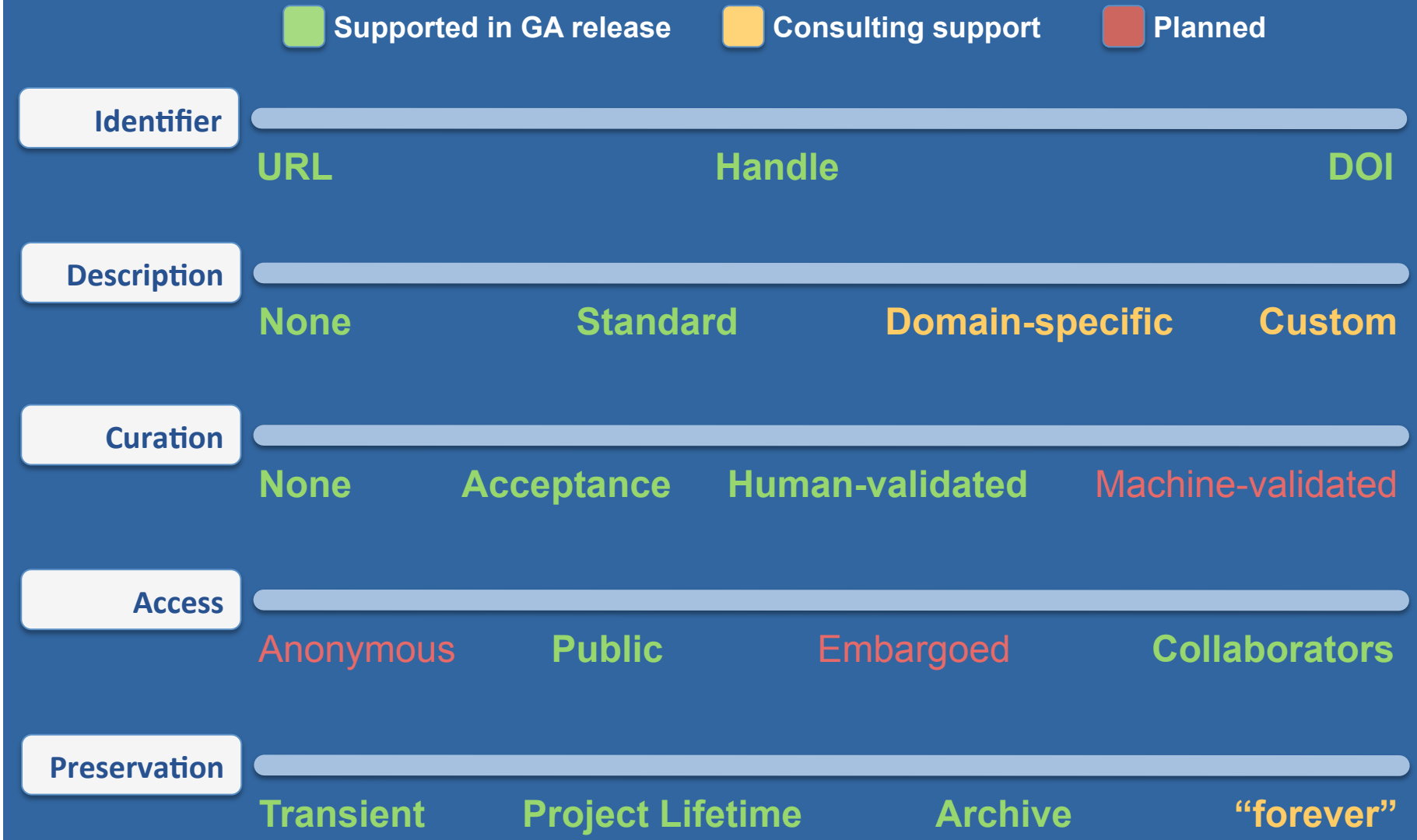
Discover

- **Announced at GW14**
- **Pilots with 8 design partners**
- **Generally available next month**

Publication spans a broad scope

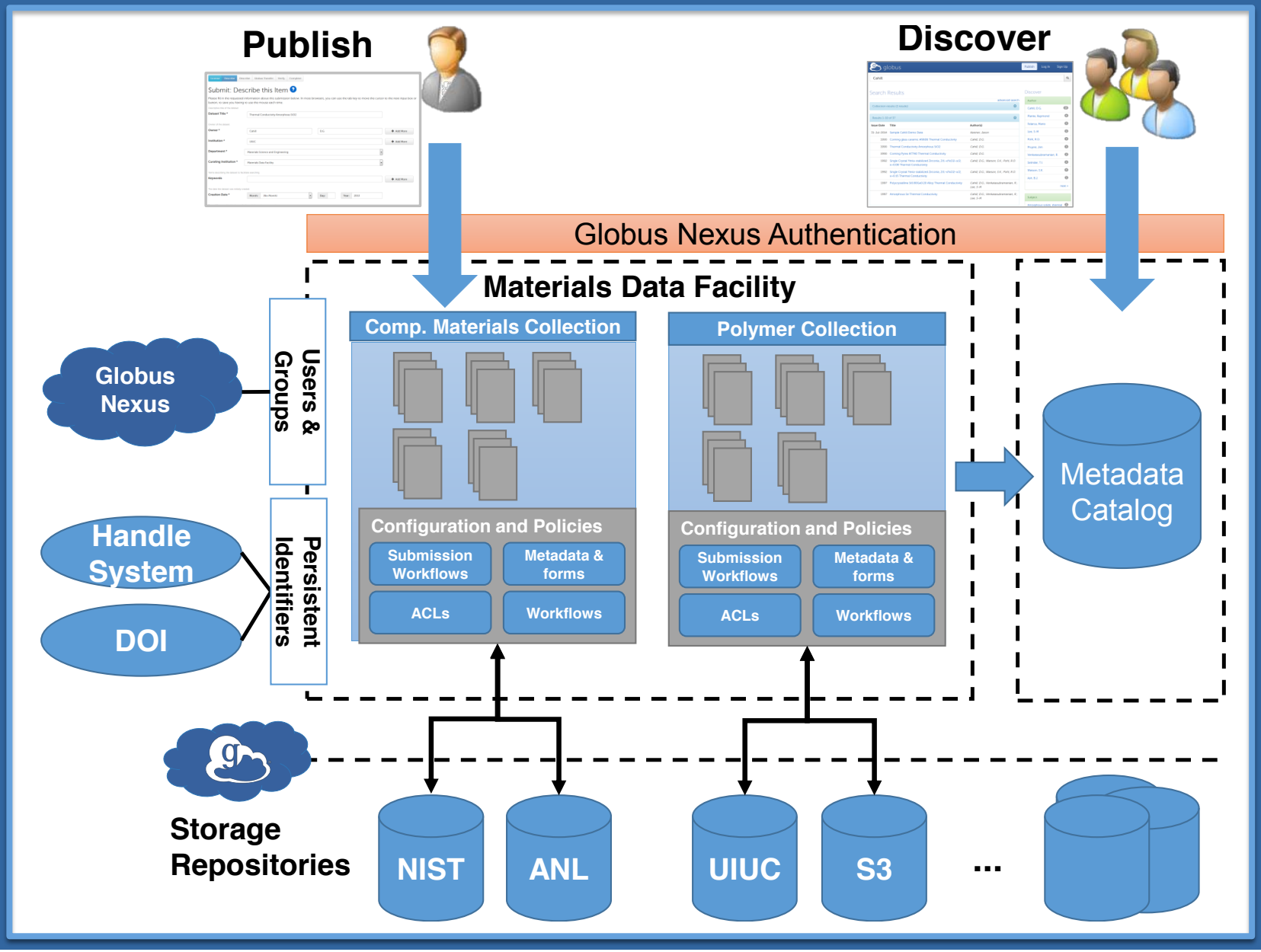


Publication GA release



Data Publication Demonstration

The Materials Data Facility



globus

Publish | Log In | Sign Up

Search

Materials Data Facility Community home page

Browse

Issue Date | Author | Title | Subject

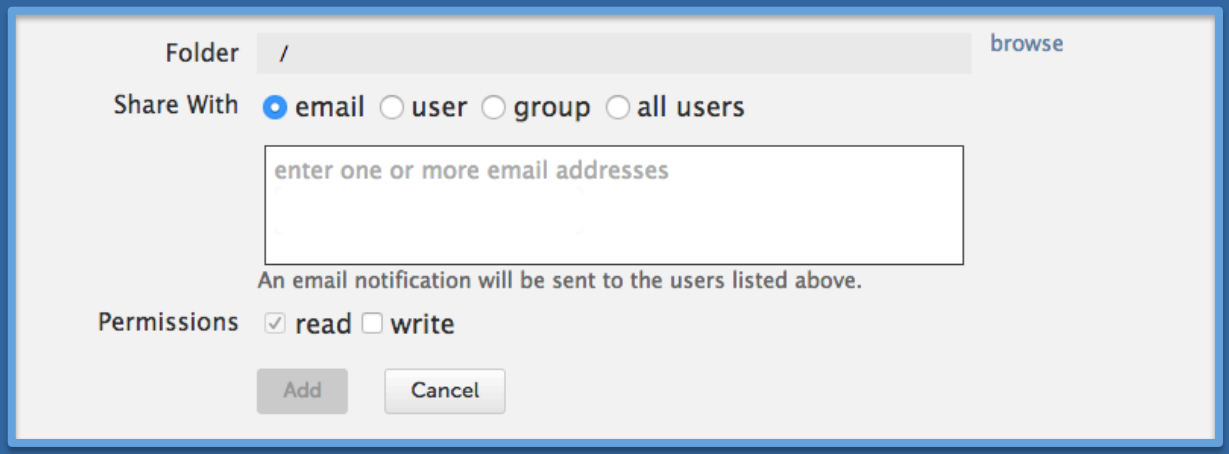
Discover

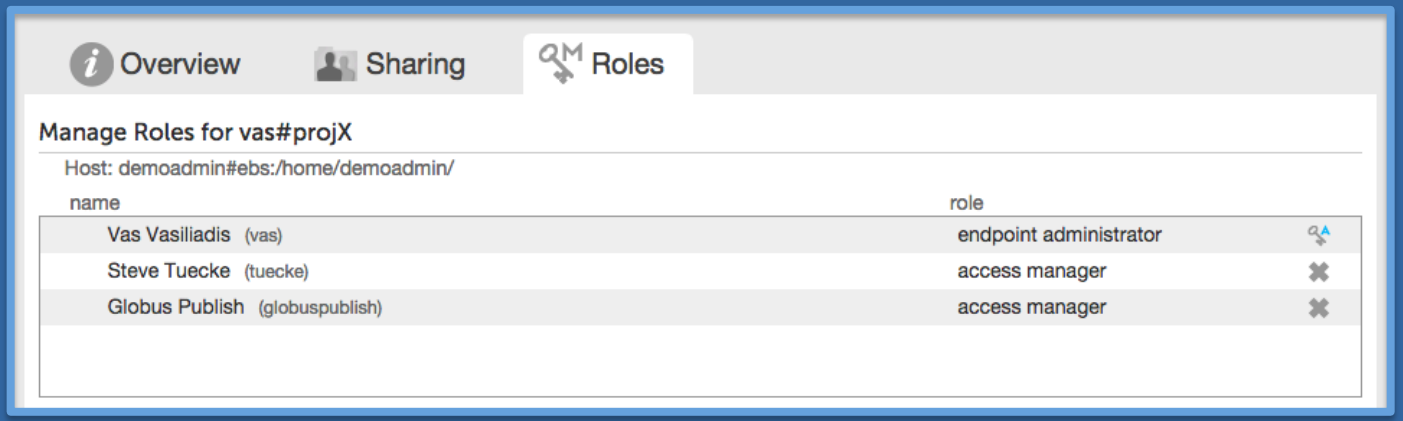
Author		Subject		Date issued	
Cahill, D.G.	12	Amorphous solids, thermal conduct...	3	2010 - 2015	14
Plante, Raymond	6	thermal conductivity, superlattic...	3	2000 - 2009	8
Felarca, Mario	5	ozone, thermal conductivity, chem...	2	1990 - 1999	13
Lee, S-M	5	thermal conductivity, single crys...	2	1987 - 1989	1
Pohl, R.O.	3	another test	1		
Pruyne, Jim	3	example test	1		
Venkatasubramanian, R.	3	Film	1		

Other data publication examples

Data sharing enhancements

- **Share to all**
- **Invite by email**
- **Access manager**

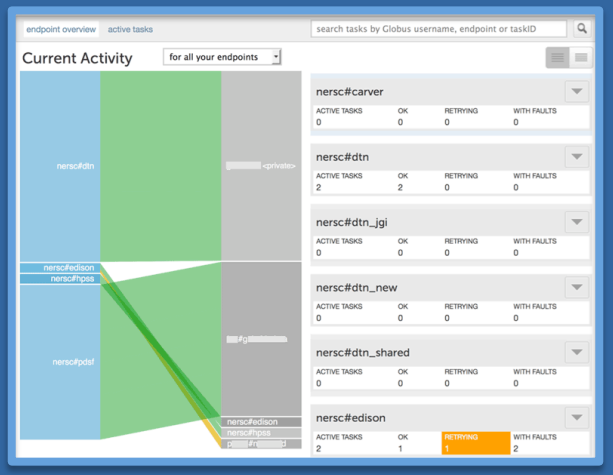




Data Sharing Enhancements Demonstration

New resource provider features

- **Management console**
  - Task monitoring
  - Cancel, pause/resume coming soon
- **Endpoint concurrency and parallelism controls**
- **Network manager plug-in**
- **Streamlined account provisioning**

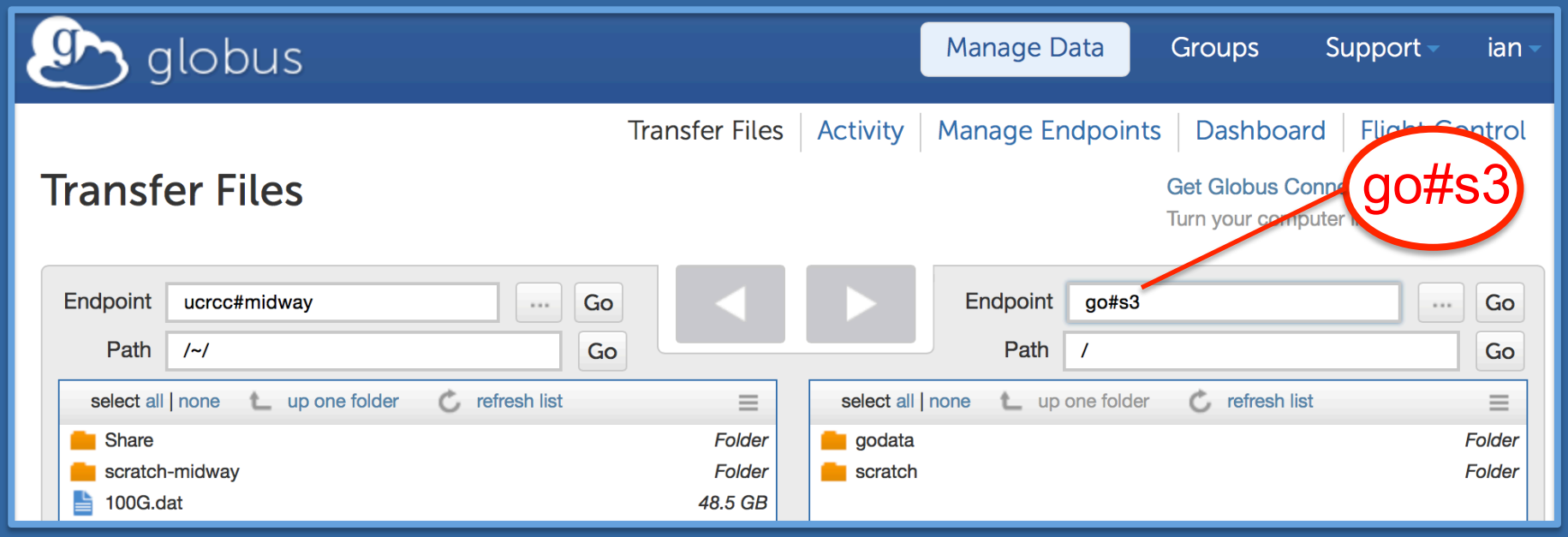


Management Console Demonstration

Bridging to public clouds

- **Amazon S3 endpoints**
  - Introduced last year, now fully supported
- **Many performance improvements**
  - Sustained multi-Gbps throughput
- **Will support OpenStack Ceph object stores, e.g. for Jetstream**

High-speed transfers to/from AWS cloud, via Globus transfer service



- **UChicago → AWS S3 (US region): Sustained 2 Gbps**
  - 2 GridFTP servers, GPFS file system at UChicago
  - Multi-part upload via 16 concurrent HTTP connections
- **AWS EC2 (ephemeral storage) → AWS S3: ~5Gbps peak**

Under-the-hood improvements

- **Identity/profile/group services migrated to PostgreSQL**
- **Ember.js framework for web application**
  - Facilitates responsive design
  - Simplifies internationalization support

Globus is SaaS …

… thus simplifying
research data management
(for both users and admins),

… on storage that you select

Globus is also PaaS…

… reducing the cost to create and maintain an ecosystem of integrated research applications,

… agnostic to programming frameworks and languages.

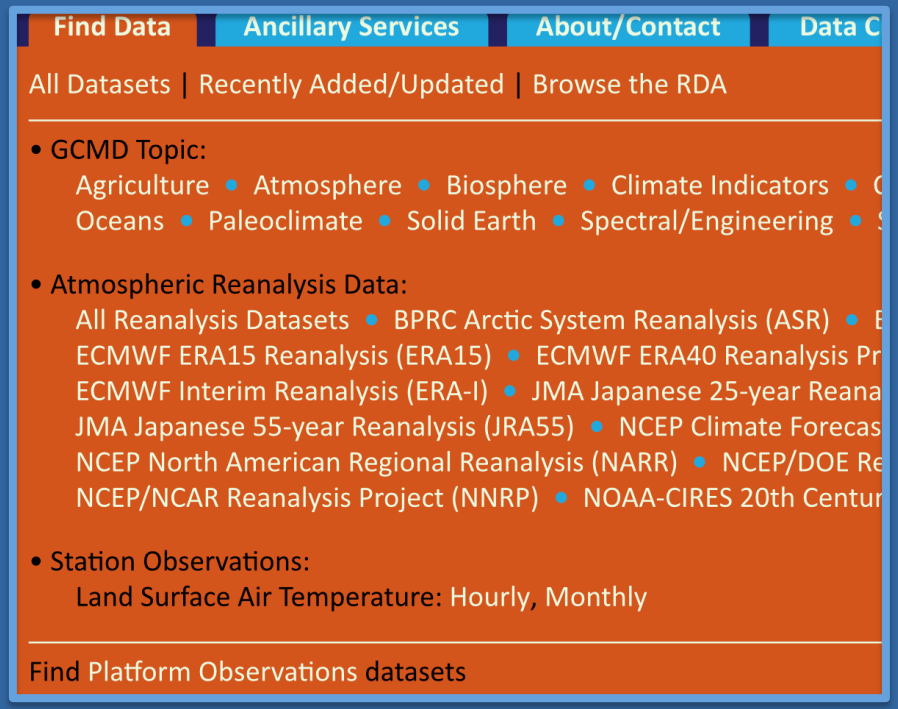
Globus Platform-as-a-Service

- **APIs initially focused on data services with some extra enabling features**
  - Transfer API
  - OAuth API
  - Group membership API
- **Streamlined account provisioning**
  - XSEDE rollout later this year

Developer web site (coming soon)

Science gateway integration

- **Example: Research Data Archive at NCAR**
- **Integrate Globus for data downloads**
- **Shared endpoint with subfolder per request**
- **Single sign on via streamlined account provisioning**



CISL Research Data Archive

Managed by NCAR's Data Support Section
Data for Atmospheric and Geosciences Research

RDA

Globus Genomics

Flexible, scalable, affordable genomics analysis for all biologists

Next-gen sequence analysis SaaS

+

Data management PaaS

+

Scalable IaaS

Galaxy

globus

amazon web services

20 groups, 2000 genomes, 1 PB, 4M CPU hours

Sustainability Update

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

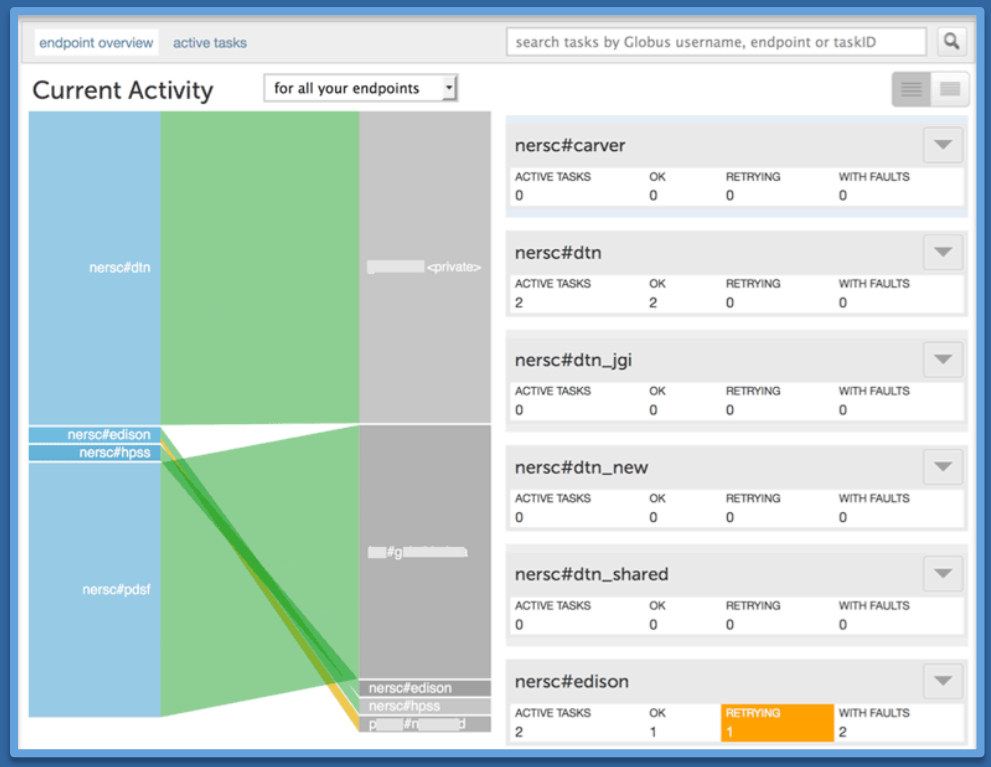
We are a non-profit, delivering a production-grade service to the non-profit research community

Our challenge:

Sustainability

Subscriptions

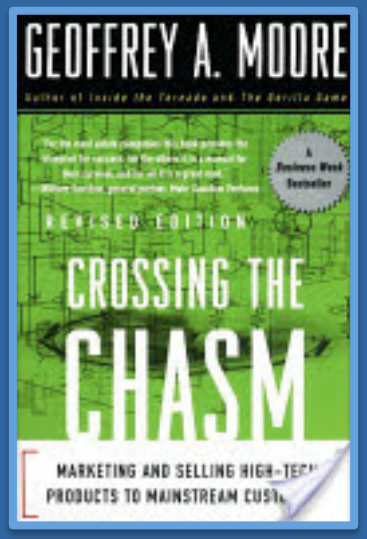
- **Globus Provider Plan**
  - Managed endpoints
  - Host shared endpoints
  - Management console
  - Data publication
  - Amazon S3 endpoints
  - Usage reporting
  - Priority support
  - Application integration

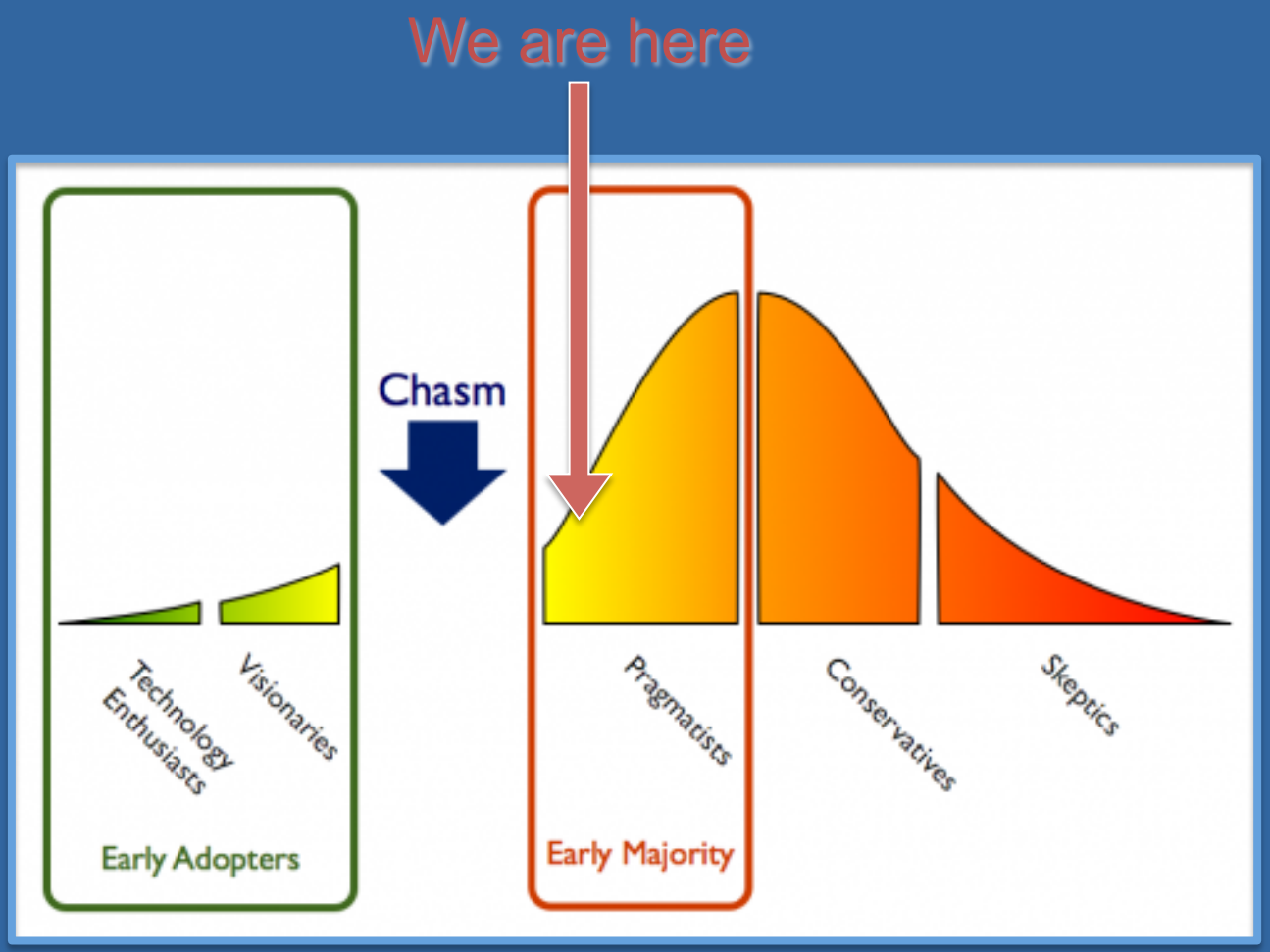


- **Branded Web Site**
- **Alternate Identity Provider (InCommon is standard)**
- **Mass Storage System optimization**

globus.org/provider-plans

Moving beyond early adopters





30+ annual subscriptions ($6k - $100k)


Serving global data
management needs

Data protection & privacy

- **Many countries have data protection laws that hinder use of US-based service**
  - E.g. EU data protection directive
- **Some US communities require higher data protection compliance**
  - HIPAA, DOE sensitive unclassified, classified

Challenges

- **Where is Globus hosted?**
  - Currently only in the US
- **Who is operating Globus?**
  - Currently a US non-profit, subject to US law
- **What are its compliance guarantees?**
  - Currently none

Federated Operation by Franchisees

- Distributed, federated operation and support of an integrated global service
- Organizations license the right to deliver Globus service to a particular community, regulatory, and legal regimes
  - UChicago provides the software
  - Franchisee provides operations, support, marketing
- Status: Building interest for initial set of franchises to fund the effort

Our vision for 21st century research data management

Provide
affordable, **advanced** capabilities
for **all** researchers,
delivering **sustainable** services
that **aggregate** and **federate**
existing resources

Thank you to our sponsors!

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY

Thank you to our partners

- **Our 25,000 registered users**
- **The many campuses who provide Globus to their users**
- **Developers of innovative applications that leverage Globus PaaS**
- **Our subscribers**
- **You!**

Enter the Counter Contest

www.globus.org/100PB

Program Preview

- **Today**
  - Lightning talks
  - Dinner reception and poster viewing
- **Tomorrow**
  - Lightning talks
  - Roundtable discussions

#globus15

@globusonline