

Managing data throughout the research lifecycle using Globus

GlobusWorld 2015

Keynote



globus



#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

Transfer

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



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

5

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

Publish

6

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

8

Discover



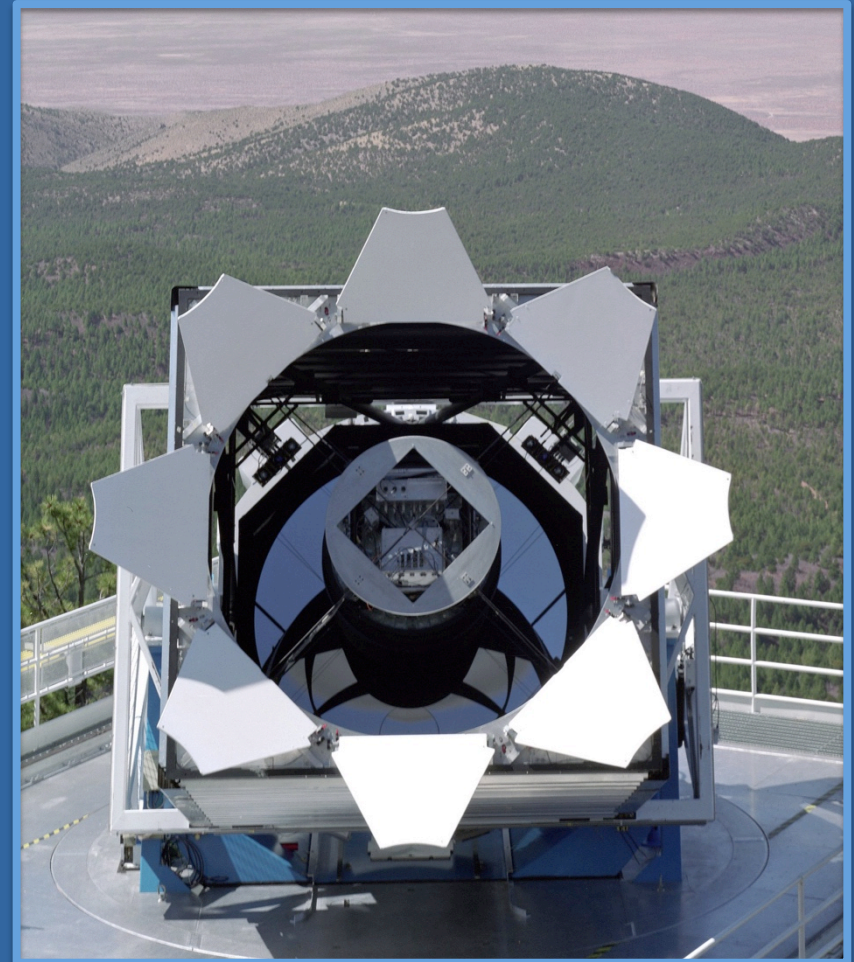
- **SaaS → Only a web browser required**
- **Access using your campus credentials**
- **Globus monitors and informs throughout**



Personal Computer



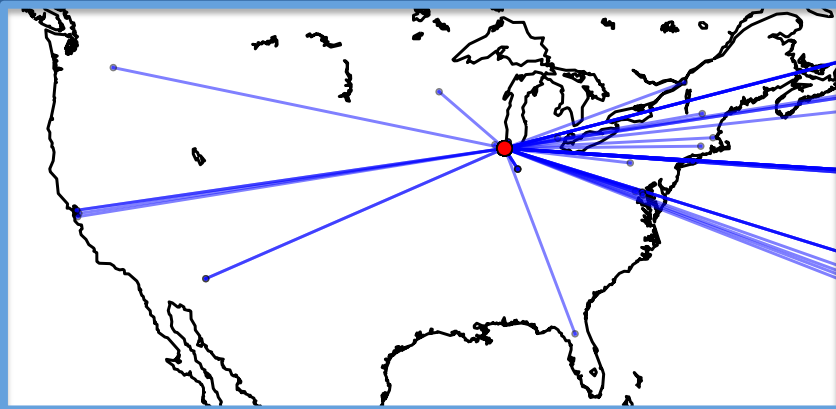
400,000 transfers
during Jan-Feb
2015 from SDSS
repository at NYU



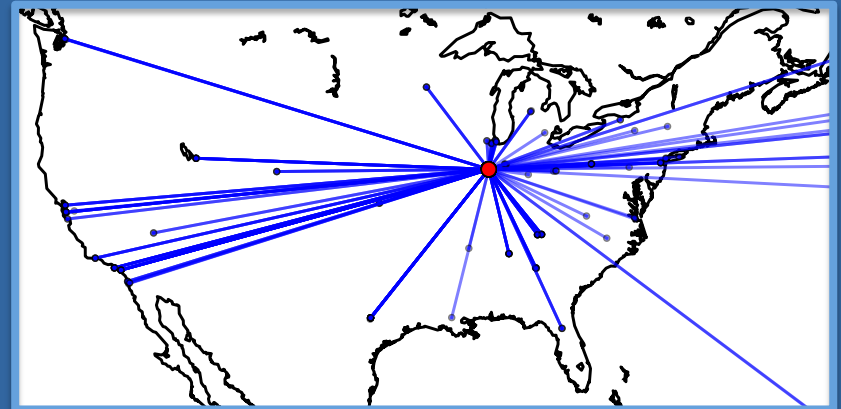
Sloan Digital Sky Survey
Source: University of Utah



People are moving data worldwide



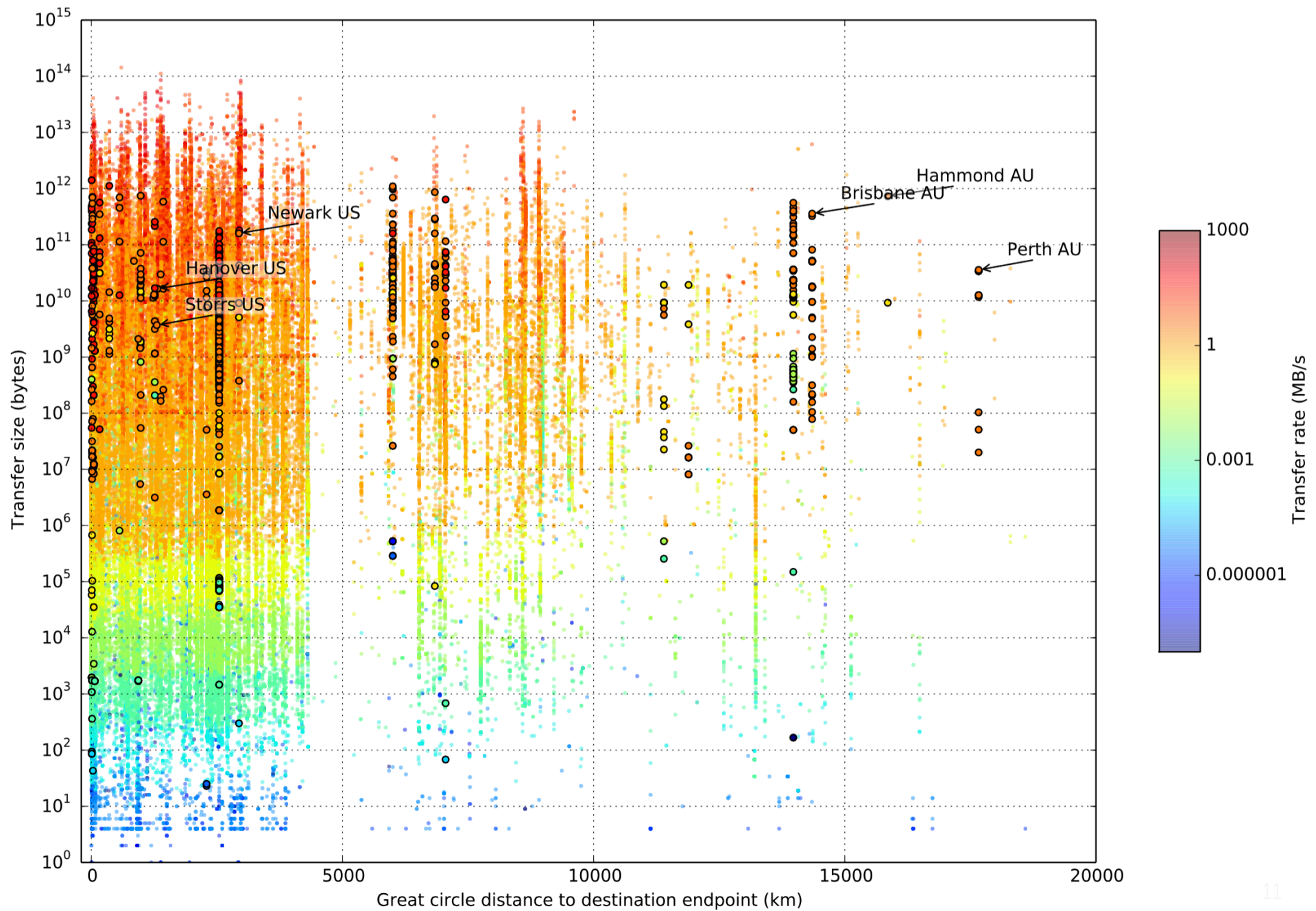
Advanced Photon Source
2,275 transfers to 119 destinations



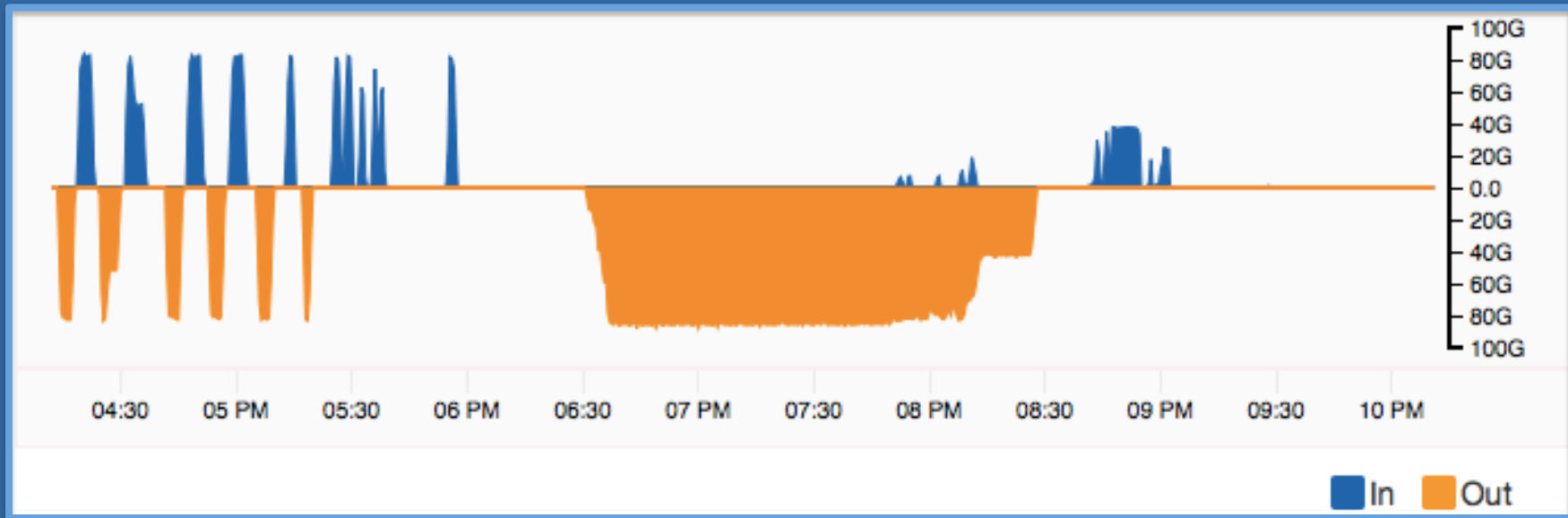
Blue Waters at NCSA
79,525 transfers to 300 destinations



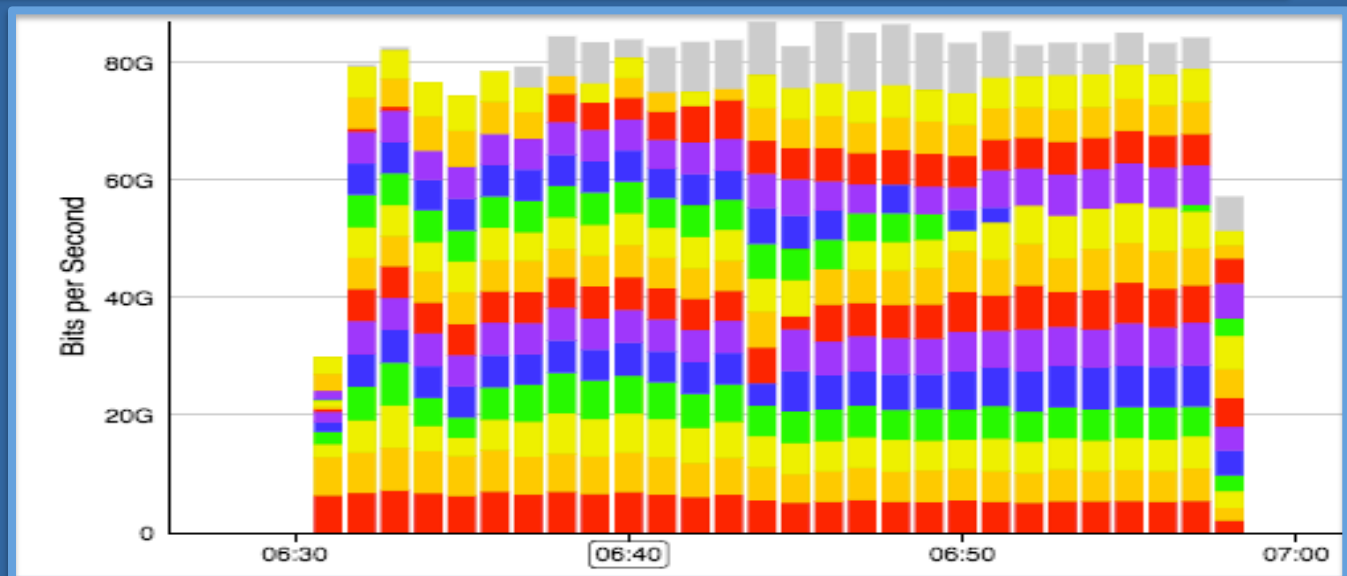
Globus transfers, showing rate (via color) as a function of distance and volume.
The 1921 transfers from aps#clutch are highlighted.



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

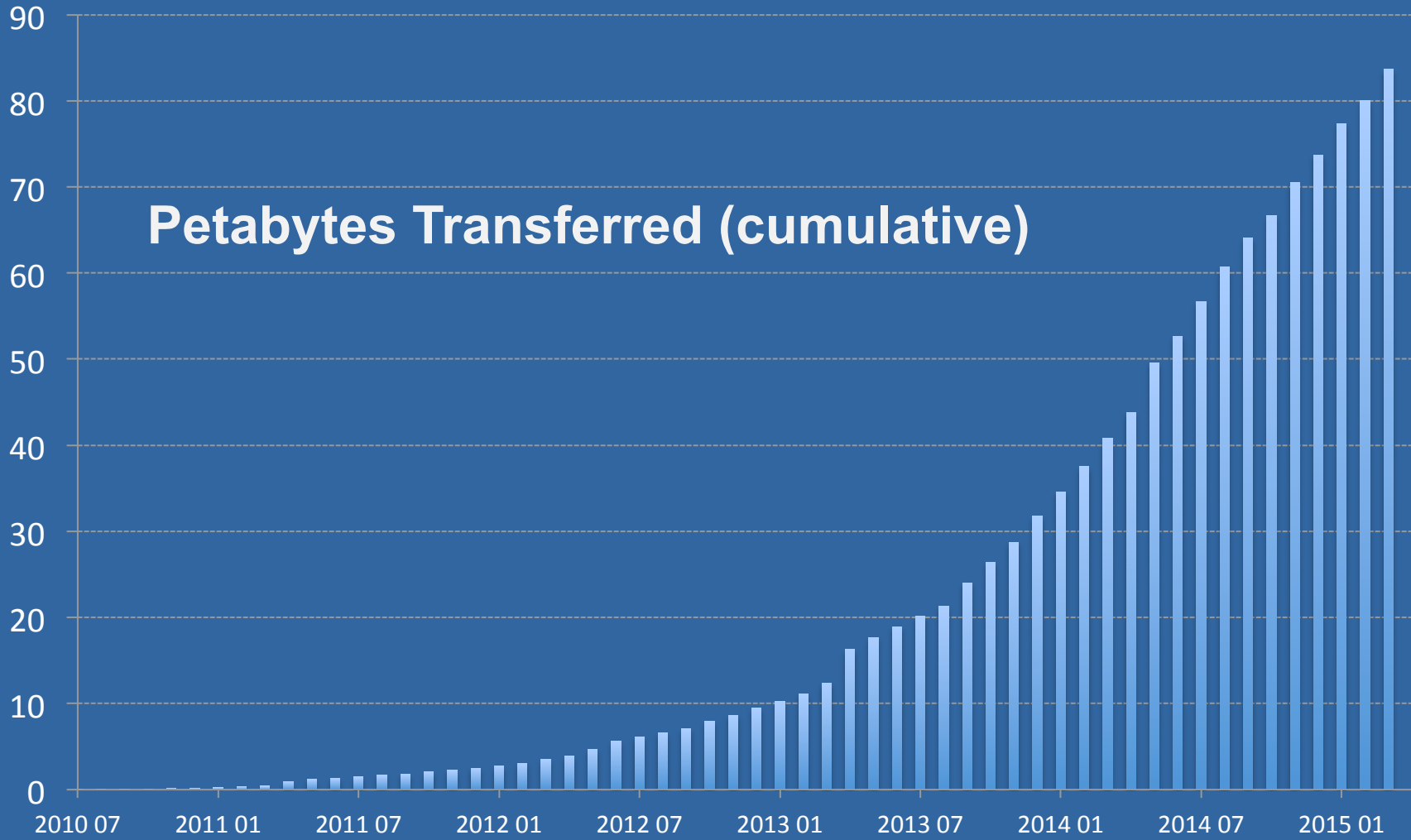


Raj Kettimuthu
and team,
Argonne
Nov 2014



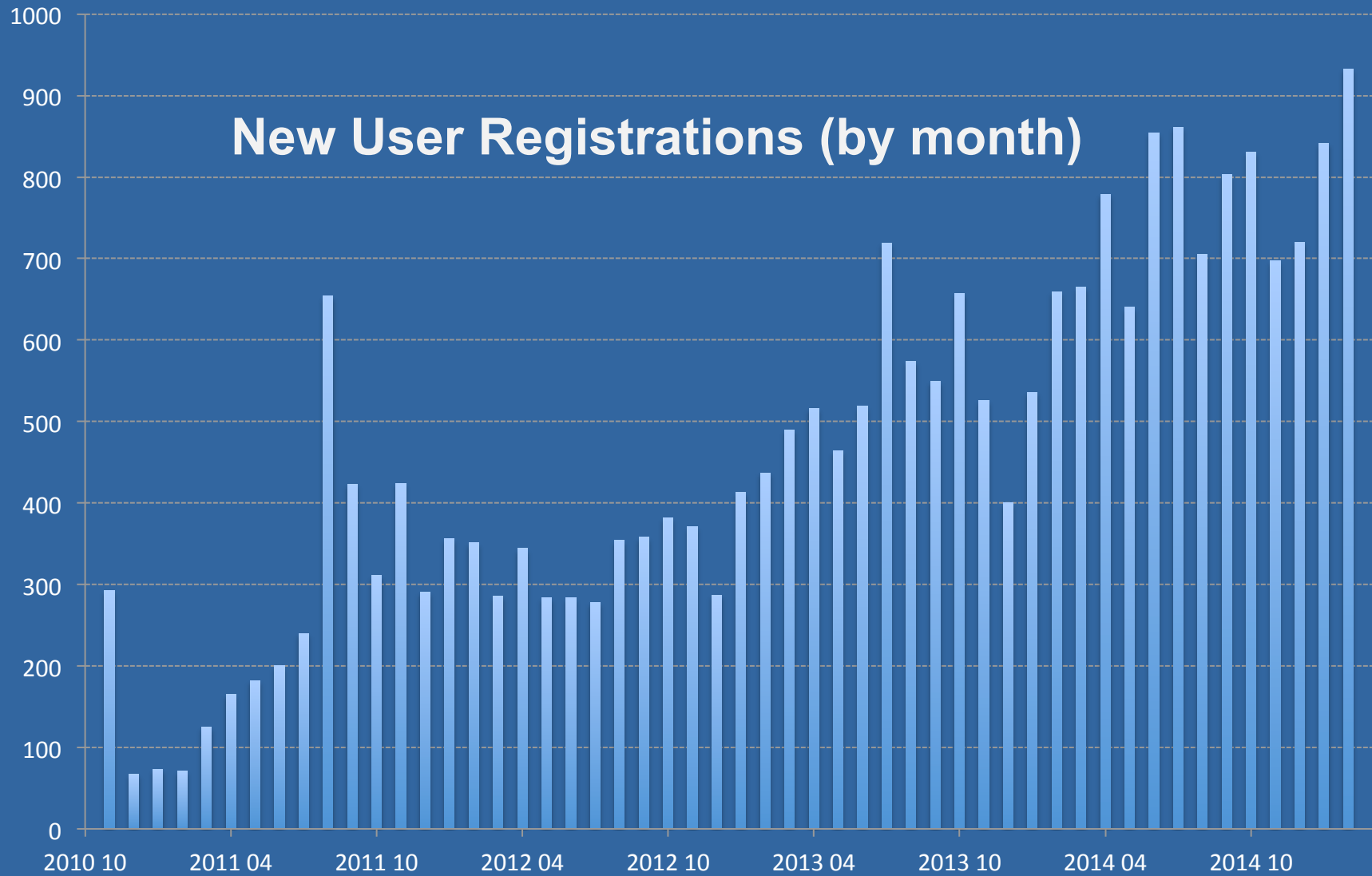


Broad adoption continues



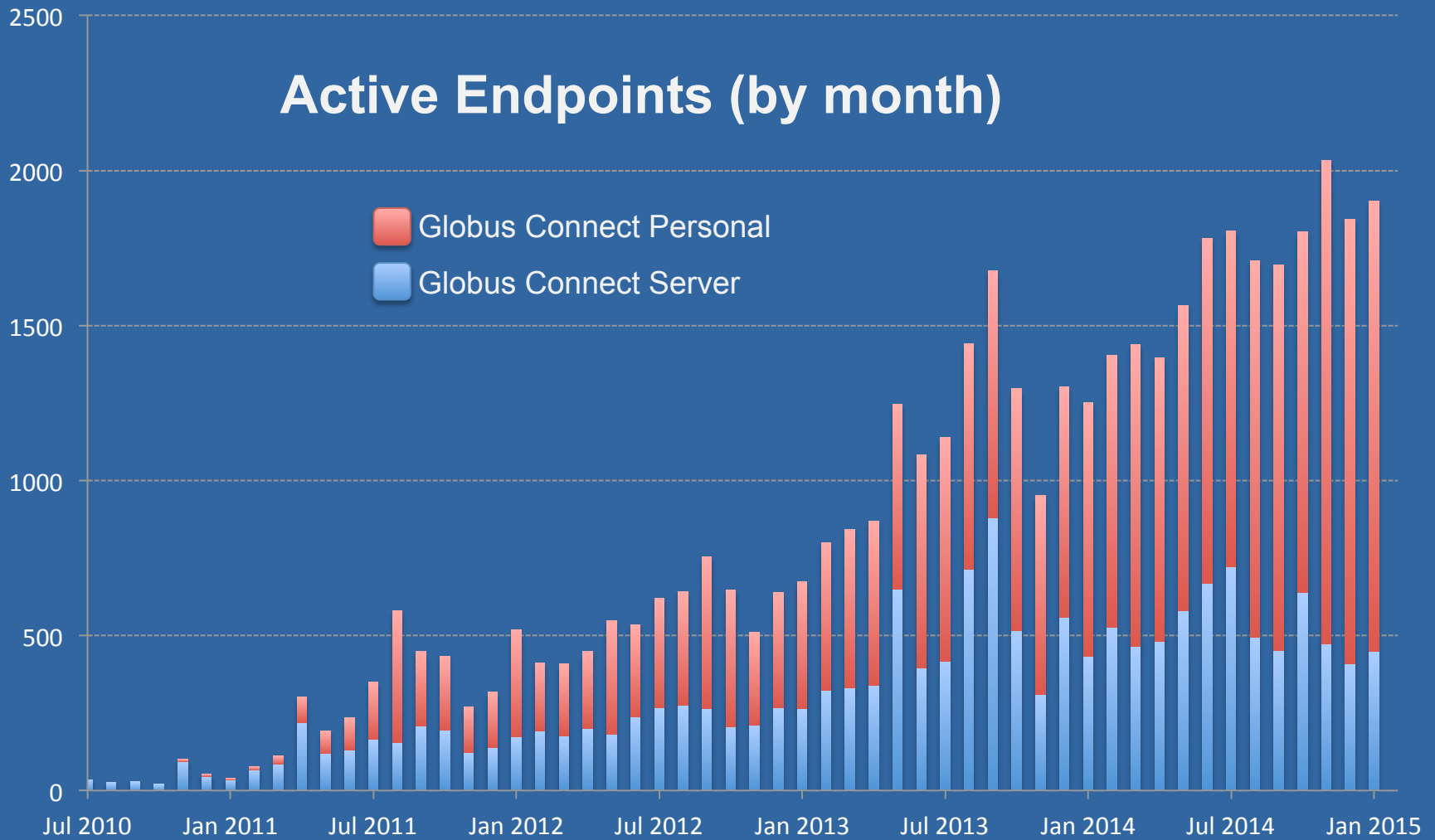


Broad adoption continues





Broad adoption continues





Data publication update

Light Source



Globus transfers files reliably, securely

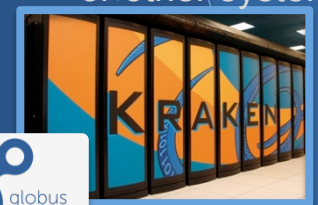
2

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



Publication Repository

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

1



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

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

5

Publish

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

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

Discover



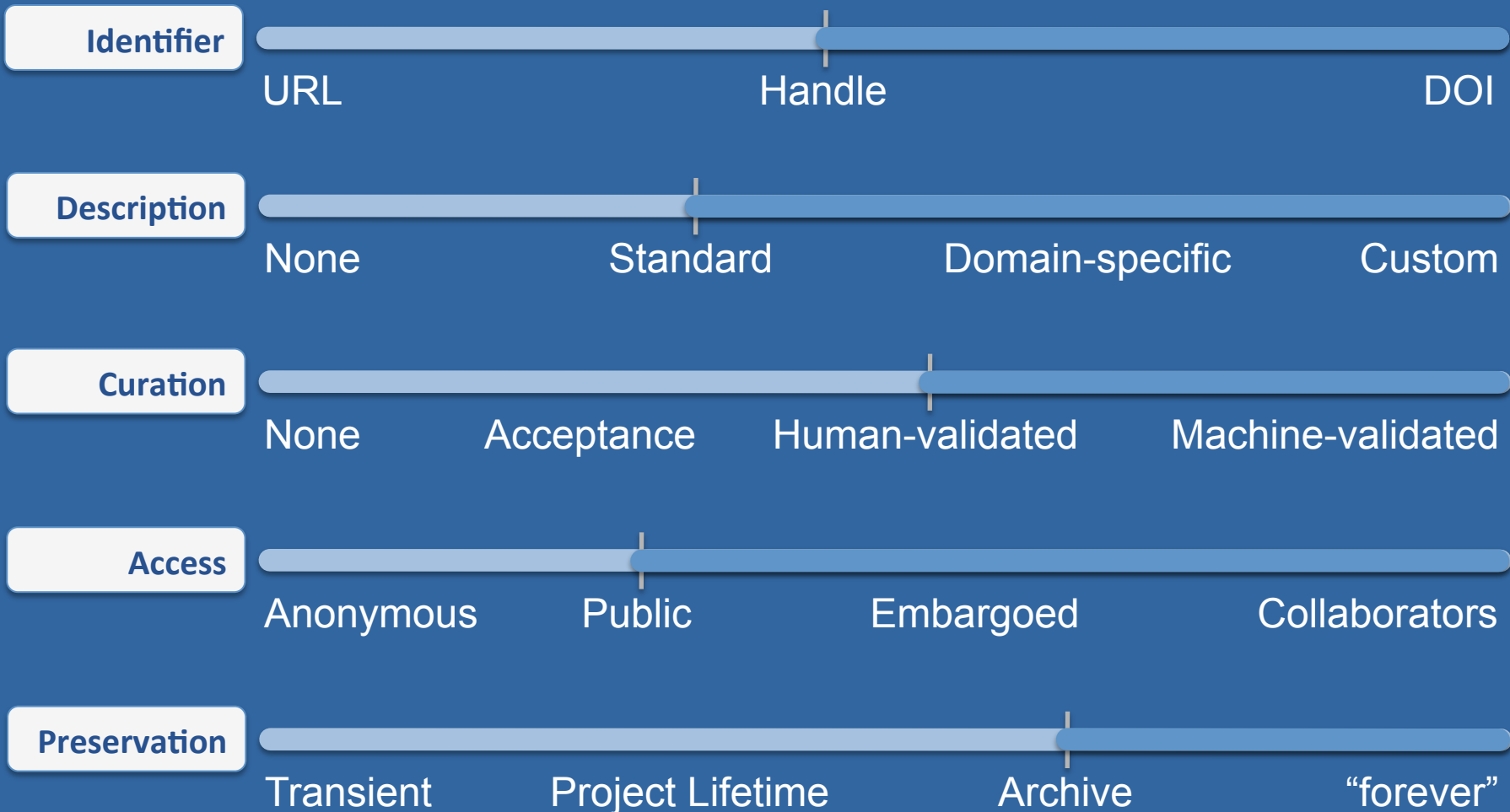
- Announced at GW14
- Pilots with 8 design partners
- Generally available next month



Personal Computer



Publication spans a broad scope





Publication GA release

■ Supported in GA release ■ Consulting support ■ Planned

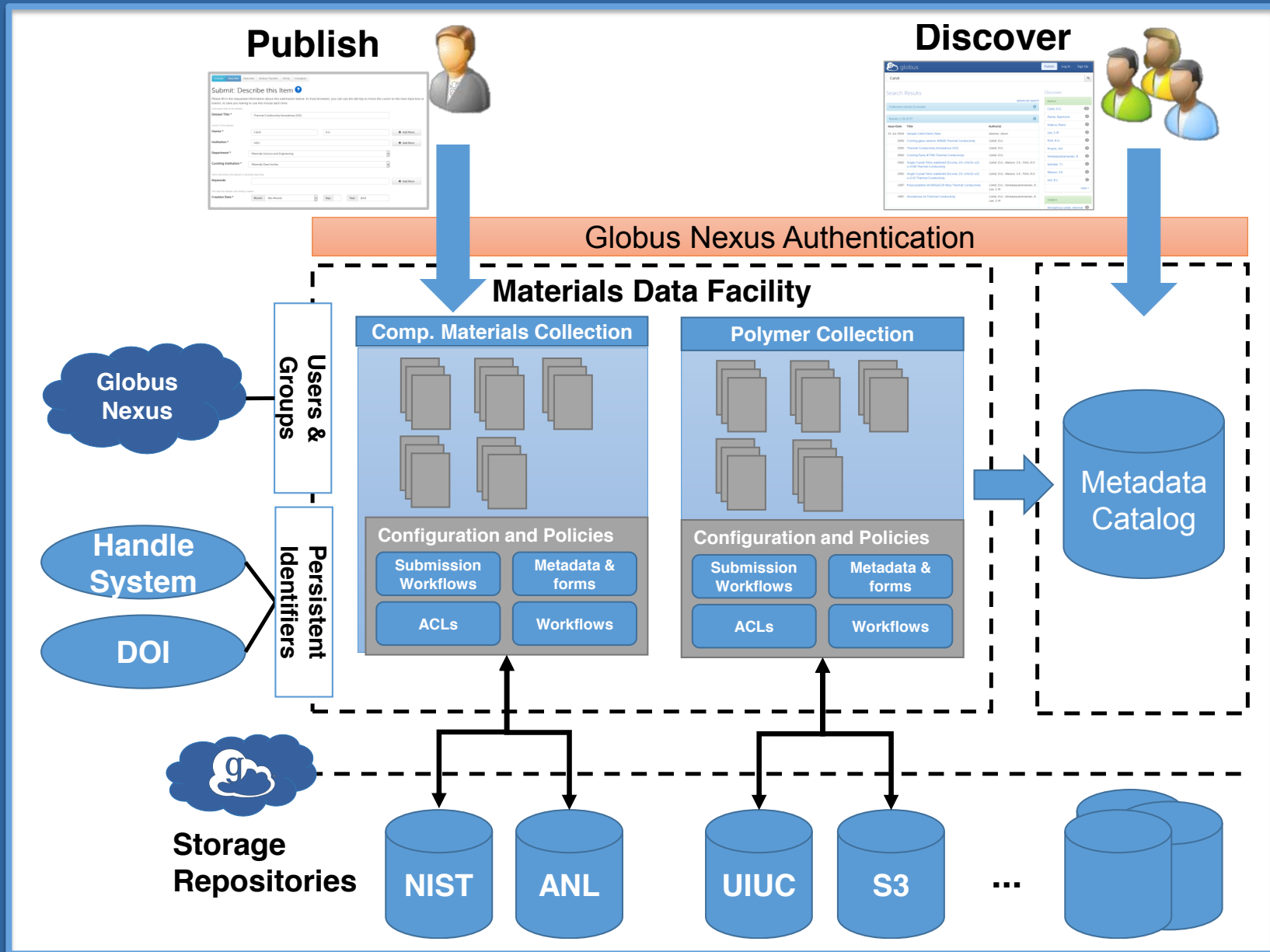
Identifier	_____			
	URL	Handle		DOI
Description	_____			
	None	Standard	Domain-specific	Custom
Curation	_____			
	None	Acceptance	Human-validated	Machine-validated
Access	_____			
	Anonymous	Public	Embargoed	Collaborators
Preservation	_____			
	Transient	Project Lifetime	Archive	“forever”



Data Publication Demonstration



The Materials Data Facility



Search

Materials Data Facility Community home page

Browse

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

Subject

Amorphous solids, thermal conduct...	3
thermal conductivity, superlattic...	3
ozone, thermal conductivity, chem...	2
thermal conductivity, single crys...	2
another test	1
example test	1
Film	1

Date issued

2010 - 2015	14
2000 - 2009	8
1990 - 1999	13
1987 - 1989	1



Other data publication examples



NEW YORK UNIVERSITY



compute  calcul
C A N A D A





Data sharing enhancements

- Share to all
- Invite by email
- Access manager

Folder [browse](#)

Share With email user group all users

An email notification will be sent to the users listed above.

Permissions read write

[i Overview](#) [Sharing](#) [QM Roles](#)

Manage Roles for vas#projX

Host: demoadmin#ebs:/home/demoadmin/

name	role	
Vas Vasiliadis (vas)	endpoint administrator	
Steve Tuecke (tuecke)	access manager	
Globus Publish (globuspublish)	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**

Endpoint	Active Tasks	OK	Retrying	With Faults
nersc#carver	0	0	0	0
nersc#dtn	2	2	0	0
nersc#dtn_jgi	0	0	0	0
nersc#dtn_new	0	0	0	0
nersc#dtn_shared	0	0	0	0
nersc#edison	2	1	0	2





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

The screenshot shows the Globus Transfer Files interface. At the top, there's a navigation bar with 'Manage Data', 'Groups', 'Support', and 'ian'. Below that, there are tabs for 'Transfer Files', 'Activity', 'Manage Endpoints', 'Dashboard', and 'Flight Control'. The main heading is 'Transfer Files'. On the right, there's a link 'Get Globus Conn' and a note 'Turn your computer'. A red circle highlights the text 'go#s3' in the top right, with an arrow pointing to the 'go#s3' endpoint input field in the right-hand pane. The left-hand pane shows an endpoint 'ucrc#midway' and a path '/~/'. The right-hand pane shows an endpoint 'go#s3' and a path '/'. Both panes have file lists with folders and files.

- **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)

globus developer

Transfer API Guides Toolkit Support

Harness the power of the Globus research data management cloud.

Transfer API Resource Providers Toolkit



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**

The screenshot shows the RDA website interface with the following elements:

- Navigation Tabs:** Find Data, Ancillary Services, About/Contact, Data C
- Search Options:** All Datasets | Recently Added/Updated | Browse the RDA
- GCMD Topic:** Agriculture • Atmosphere • Biosphere • Climate Indicators • Oceans • Paleoclimate • Solid Earth • Spectral/Engineering • S
- Atmospheric Reanalysis Data:** All Reanalysis Datasets • BPRC Arctic System Reanalysis (ASR) • ECMWF ERA15 Reanalysis (ERA15) • ECMWF ERA40 Reanalysis Pr • ECMWF Interim Reanalysis (ERA-I) • JMA Japanese 25-year Reana • JMA Japanese 55-year Reanalysis (JRA55) • NCEP Climate Forecas • NCEP North American Regional Reanalysis (NARR) • NCEP/DOE Re • NCEP/NCAR Reanalysis Project (NNRP) • NOAA-CIRES 20th Centur
- Station Observations:** Land Surface Air Temperature: Hourly, Monthly
- Footer:** Find Platform Observations datasets

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



globus

+

Scalable IaaS



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

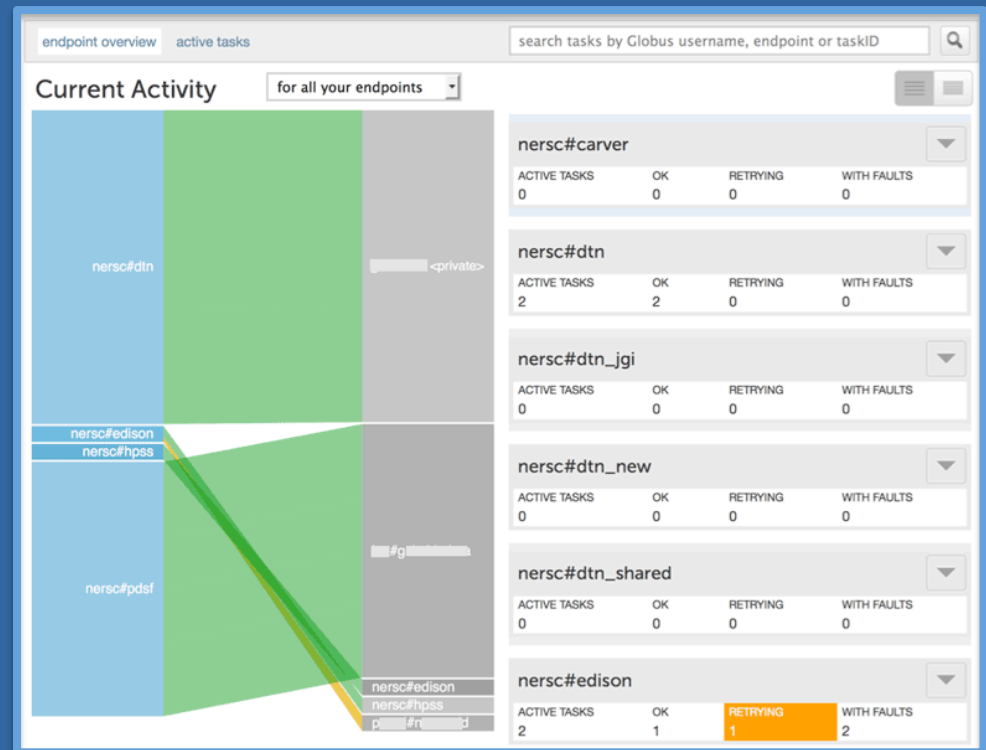
- **Globus Provider Plan**

- 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)**

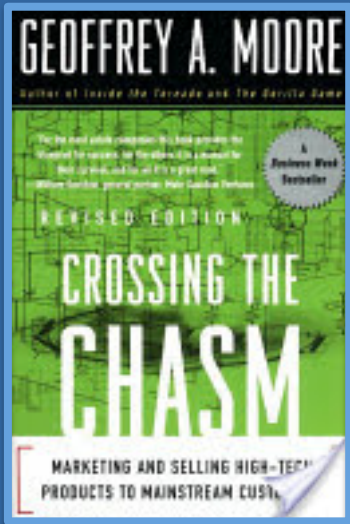
- **Mass Storage System optimization**



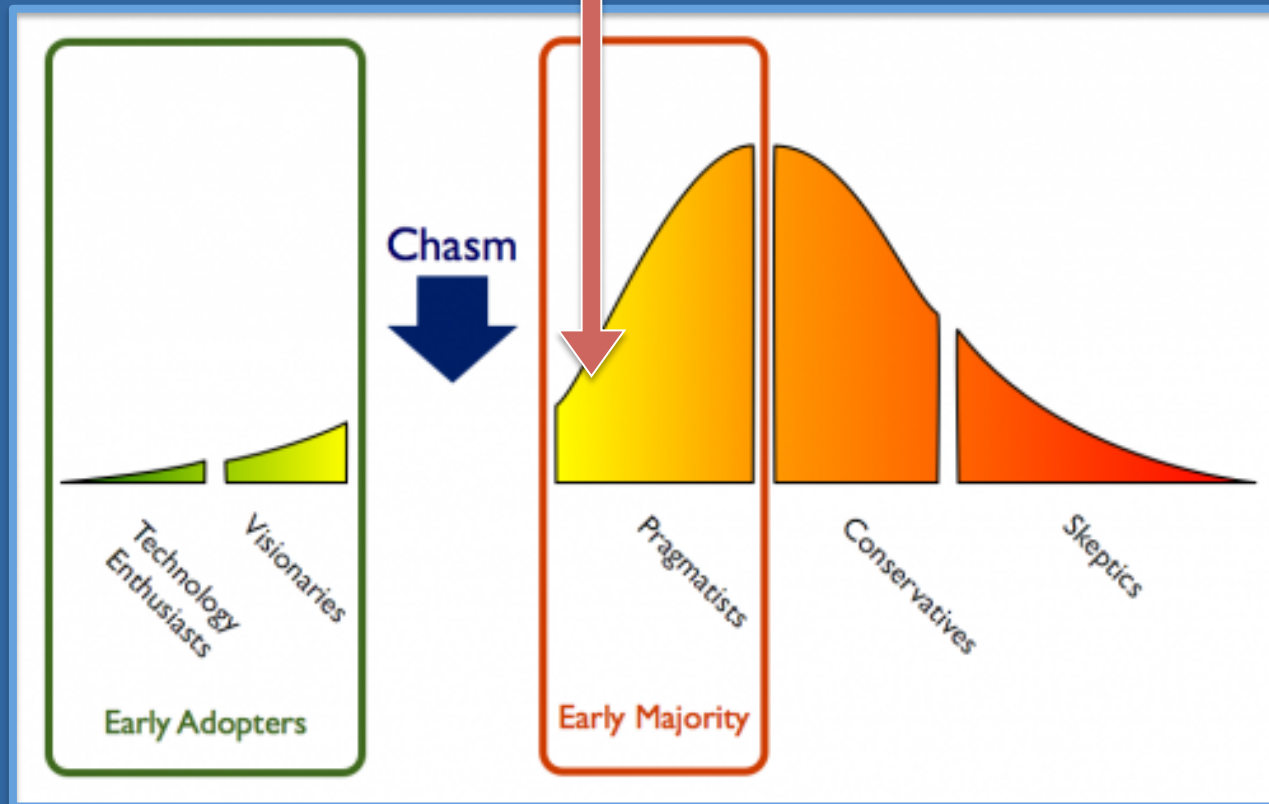
globus.org/provider-plans



Moving beyond early adopters



We are here



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



powered by
amazon
web services



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