globus

The Globus Journey:
Achieving sustainable research infrastructure for all

Ian Foster

GlobusWORLD 2013

Globus ecosystem evolution

2013

Software-as-a-Service
for sequencing analysis

2010

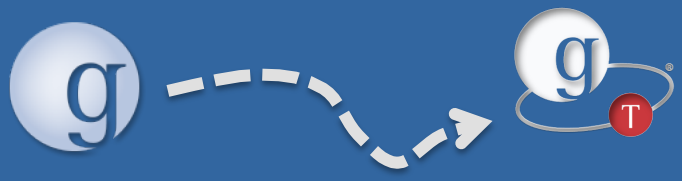
Software-as-a-Service for
Research Data Management

1998

Open source software for distributed
resource integration and access

Our vision

Accelerate discovery by individual researchers and **reduce costs** for both individuals and institutions by providing robust **research data management as a service**



We started with technology proven in many large-scale grids

GridFTP
GRAM
MyProxy
GSI-OpenSSH
…

GT usage remains strong 15 years later

3,761 GridFTP servers reporting usage

169 million GRAM jobs submitted *by the Open Science Grid in 2012*

300,000 jobs/day reported

382 million operations

29 petabytes transferred *during February 2013*

Computation Institute

THE UNIVERSITY OF CHICAGO

Argonne NATIONAL LABORATORY

Open Science Grid

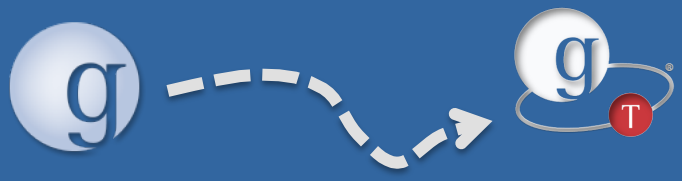
CERN

100 MyProxy servers

2,000,000 requests per week

1,000 GSI-OpenSSH servers

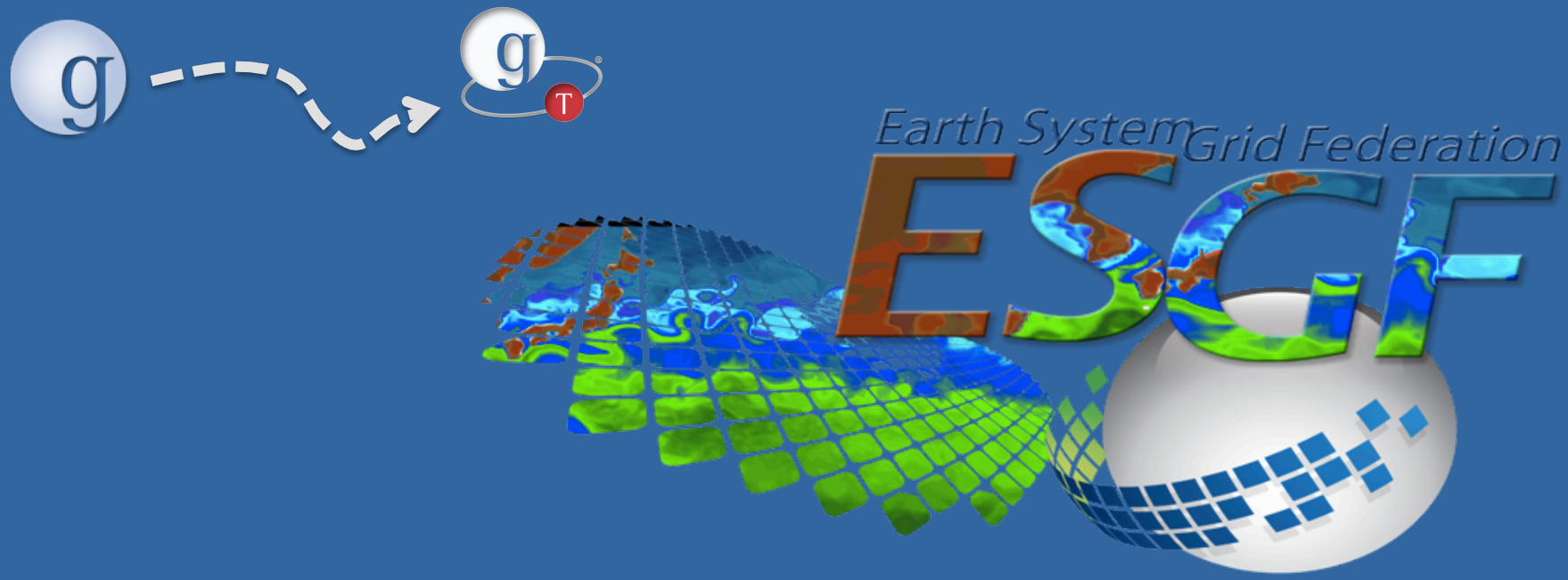
1,000,000 login requests per week



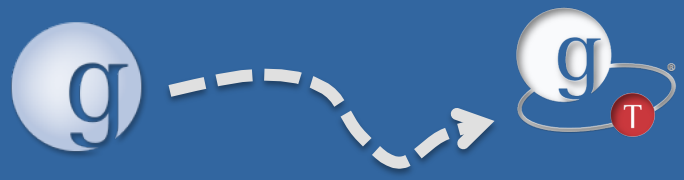
GT5.2 - 4 point releases during the past year

Focus on stability

Expand Globus Online support

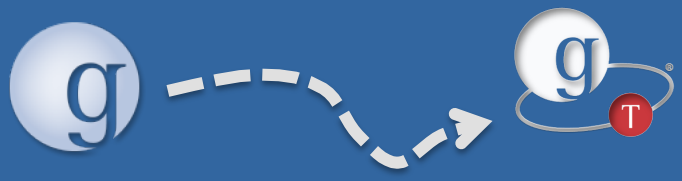


1.2 PB of climate data delivered to 23,000 users



1.2 PB of climate data delivered to 23,000 users

Typical of large, well funded research projects using GT



GT provides robust infrastructure for the 1%

GT provides robust infrastructure for the 1%

What about the 99%?

GT provides robust infrastructure for the 1%

What about the 99%?

BIG SCIENCE. Small labs

Need: A new way to deliver research cyberinfrastructure

Frictionless
Affordable
Sustainable

We asked ourselves:

What if the research work flow could be managed as easily as …

… our pictures

… our e-mail

… our entertainment

What makes these services great?

Great **User Experience**

+

Invisible, cloud-hosted infrastructure

We aspire to create a great user experience for **research data management**

We aspire to create a great user experience for **research data management**

What would a "dropbox for science" look like?

- Collect
- Move
- Sync
- Share
- Analyze
- Annotate
- Publish
- Search
- Backup
- Archive

...for BIG DATA

globus online

We adopted SaaS approaches to transform the user experience

… for both researchers and resource owners/sysadmins

Managing data should be easy …

Staging Store

Ingest Store

Registry

Analysis Store

Community Store

Archive

Mirror

… but it's hard and frustrating!

Staging Store

Archive

Mirror

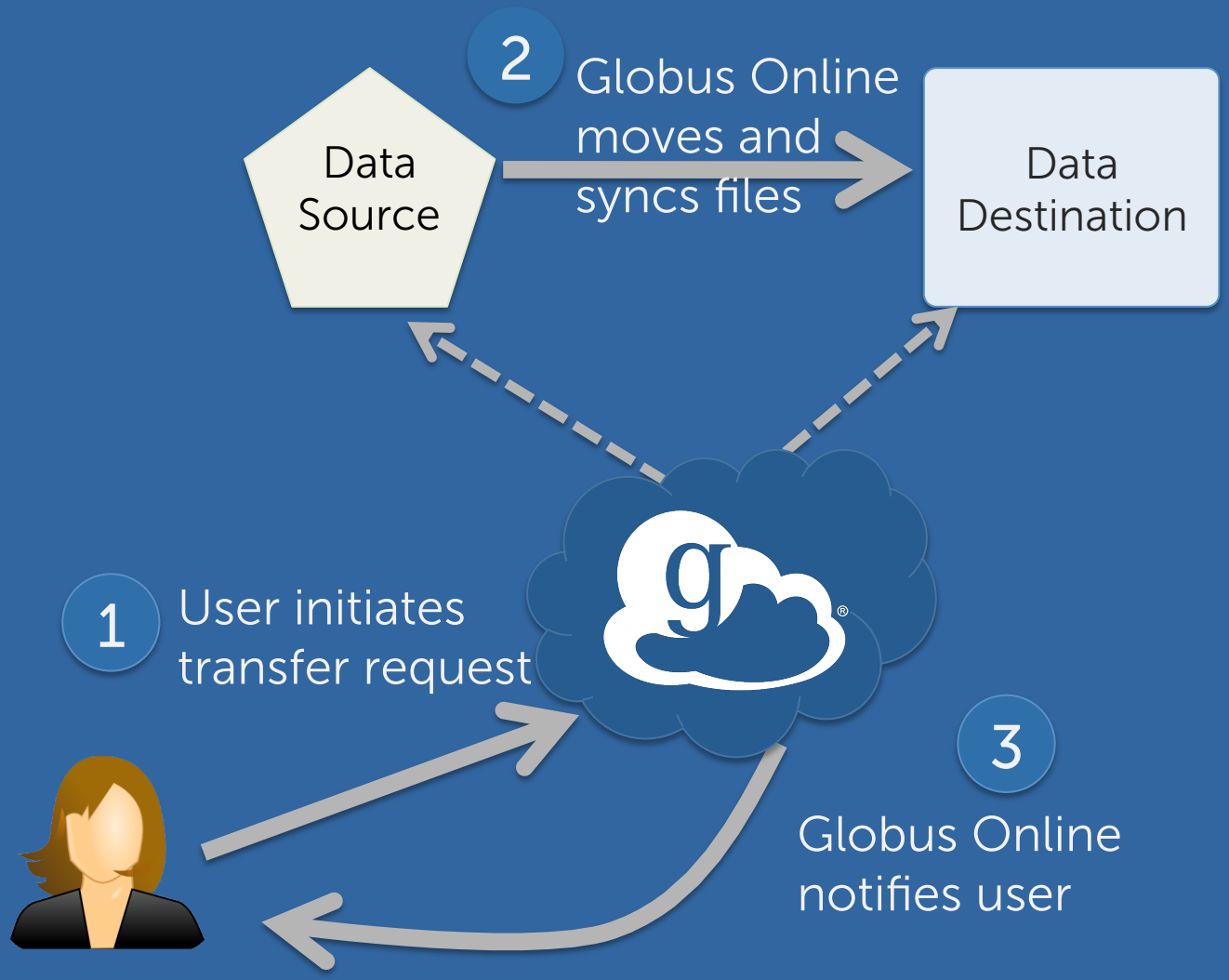
Expired credentials

Permission denied

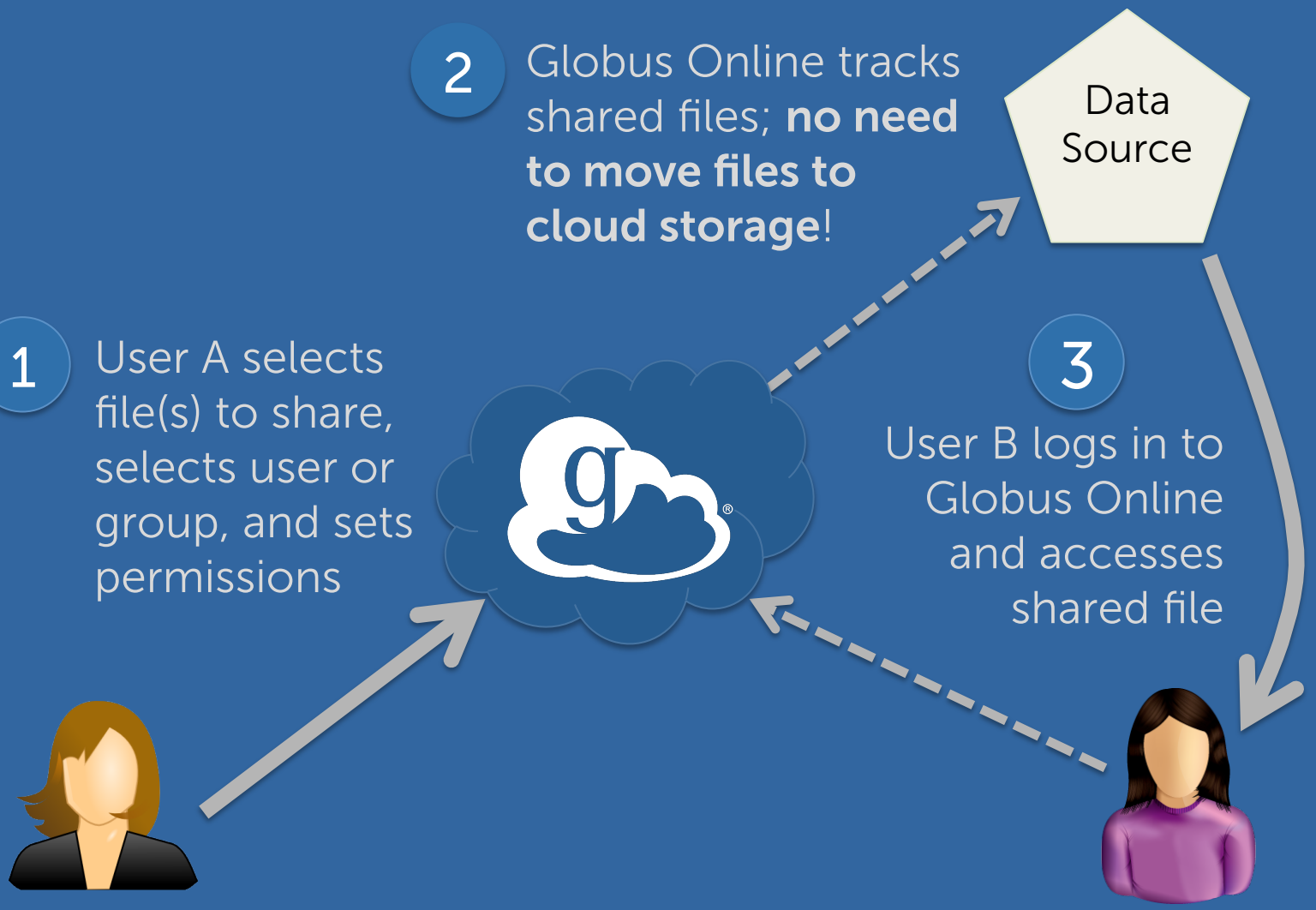
Quota exceeded

Network failed. Retry.

We started with reliable, secure, high-performance **file transfer** …

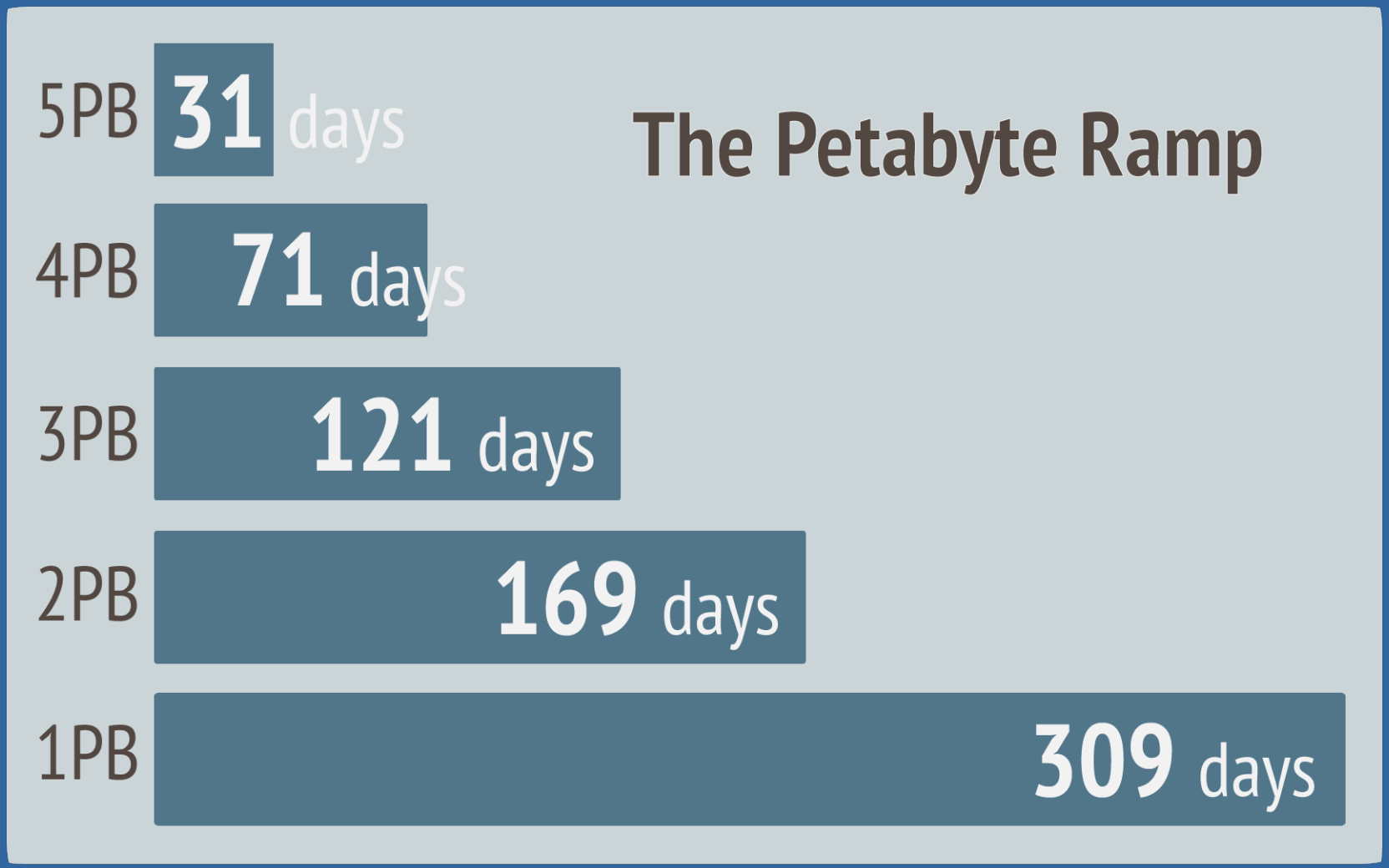


… and then made it simple to share big data off existing storage systems



Demonstration

Early adoption was slow ...

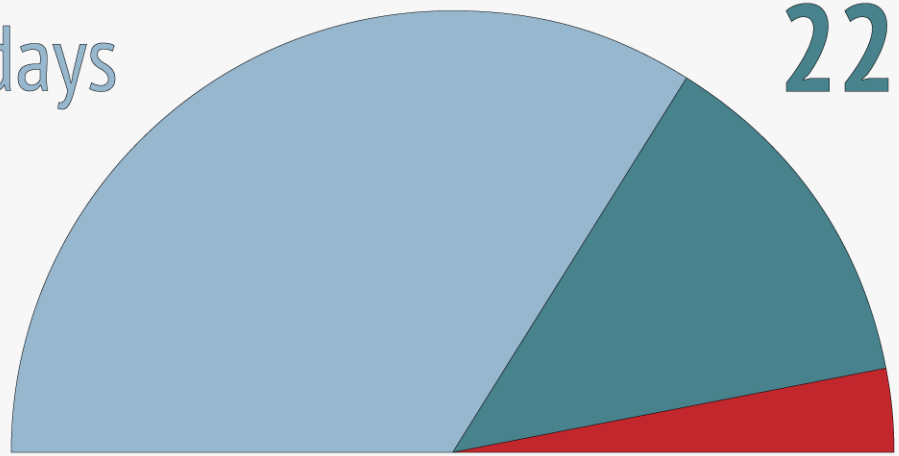


… now we're moving!

Moving the Needle

0 - 5PB in
701 days

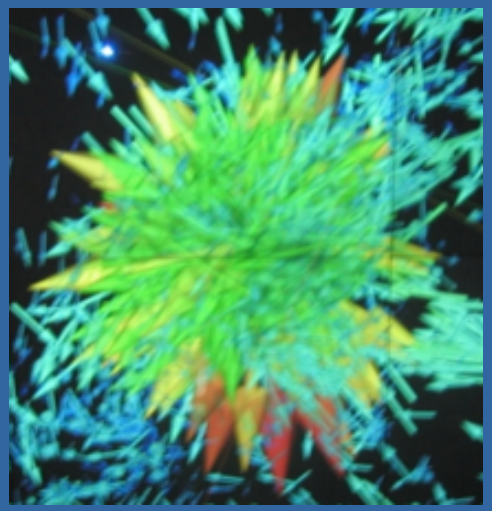
5 - 10PB
229 days



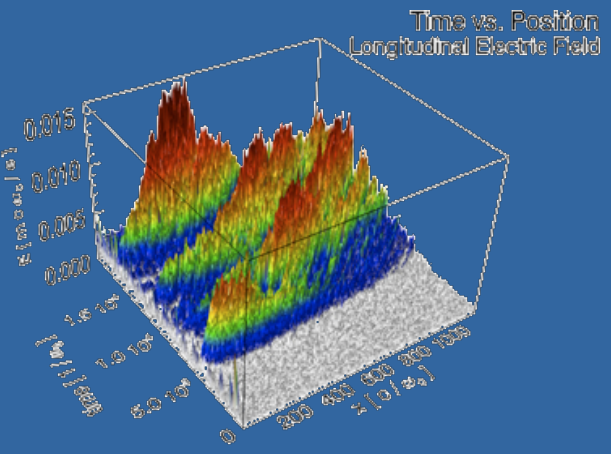
10 - 15PB
82 days

15PB moved as of April 14, 2013

9,999,988,942 MB TRANSFERRED



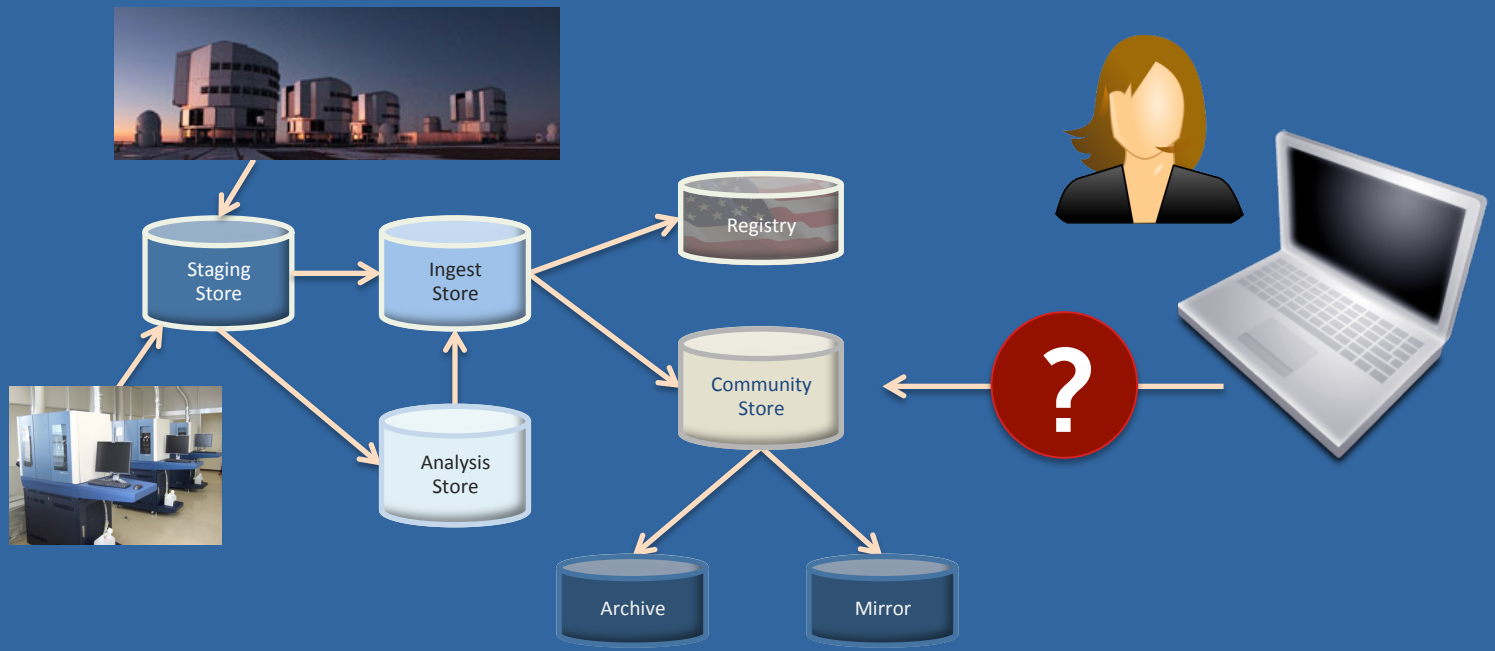
K. Heitmann (Argonne)
moves 22 TB of **cosmology** data
LANL → ANL at 5 Gb/s



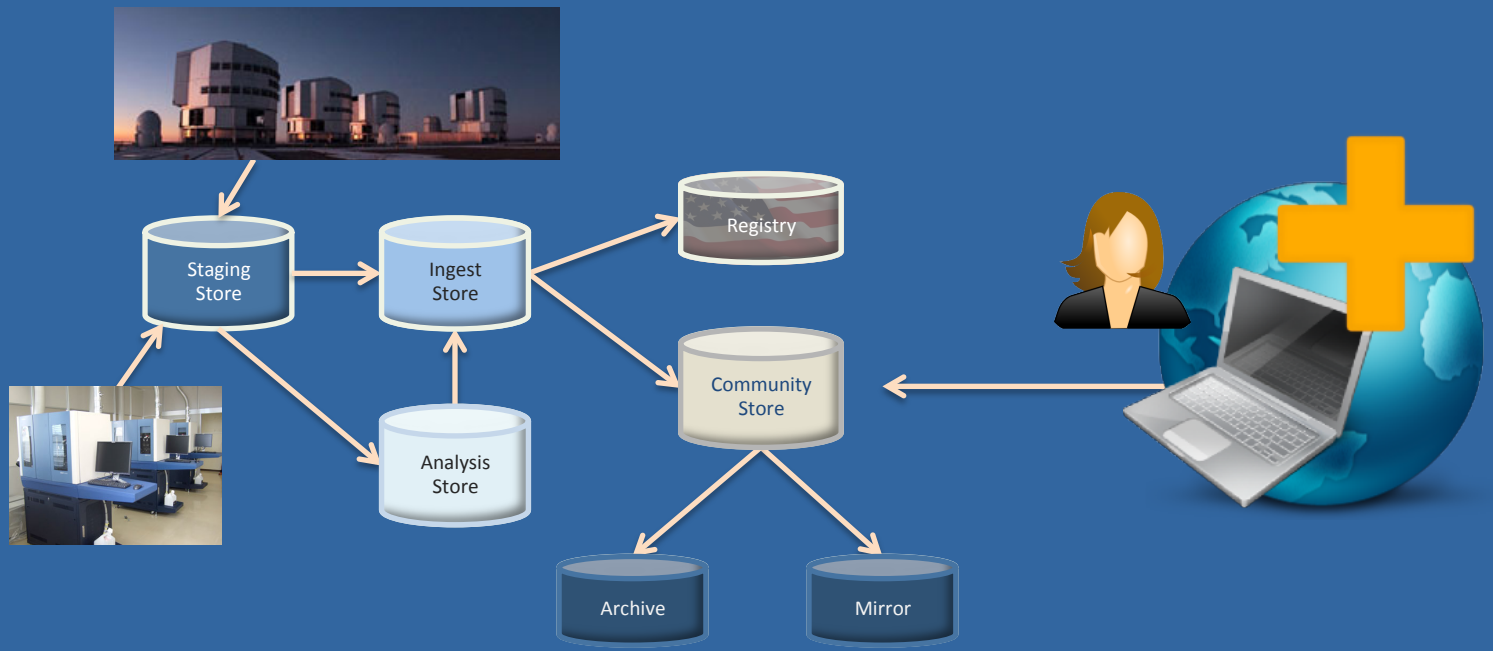
B. Winjum (UCLA)
moves 900,000-file
plasma physics datasets
UCLA →NERSC

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

The Last Mile was always the biggest challenge



Globus Connect (released Nov. 2011) enables easy connection of resources to Globus services



Globus Connect Multiuser
for resource providers

Advanced data management services to researchers

A seamlessly integrated user experience

Reduced support burden

Get started now – it's free.

globusonline.org/gcmu

Coming soon to a campus near you

XSEDE
Extreme Science and Engineering Discovery Environment

NERSC

VE RI TAS HARVARD

MICHIGAN

UNIVERSITY OF EXETER

UNIVERSITY of WASHINGTON

NCSA

Carnegie Mellon University

NGS

APS physics

NORTHWESTERN UNIVERSITY QUAECUMQUE SUNT VERA 1851

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

ISI
Information Sciences Institute

KSU

NASA

INDIANA UNIVERSITY

Fermilab

THE JOHNS HOPKINS UNIVERSITY BALTIMORE 1876

EMORY

ESnet
Energy Sciences Network

BERKELEY LAB
Lawrence Berkeley National Laboratory

Los Alamos
NATIONAL LABORATORY
EST.1943

CERN

CORNELL UNIVERSITY

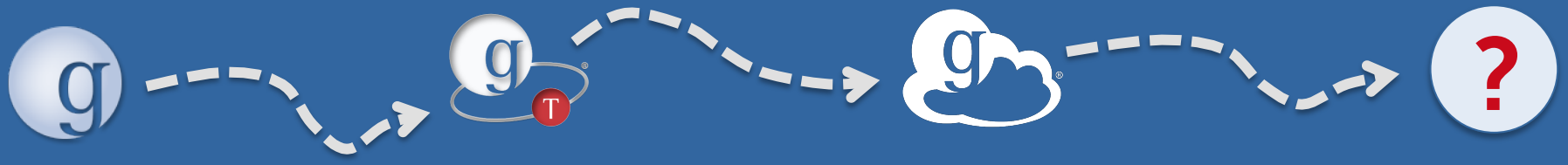
Cal

Ole Miss

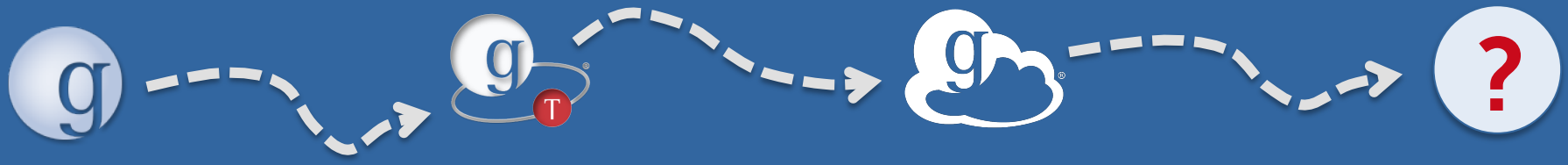
THE UNIVERSITY OF CHICAGO

NEW YORK UNIVERSITY

Argonne
NATIONAL LABORATORY



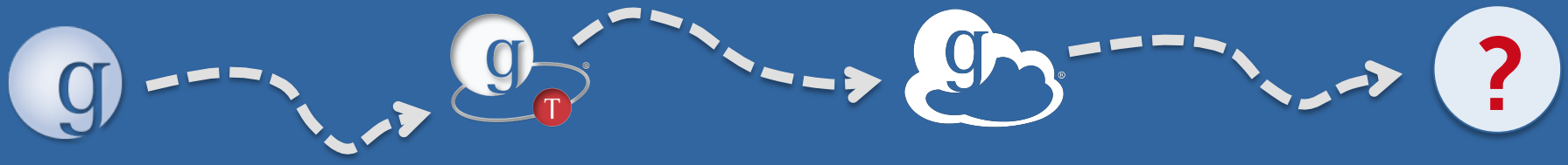
We are a non-profit service provider to the non-profit research community



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…

- Provider endpoints with sharing
- Multiple GridFTP servers per endpoint
- Branded web sites
- Alternate identity provider
- Usage reporting
- MSS optimizations
- Operations monitoring and management
- Input into and access to product roadmap

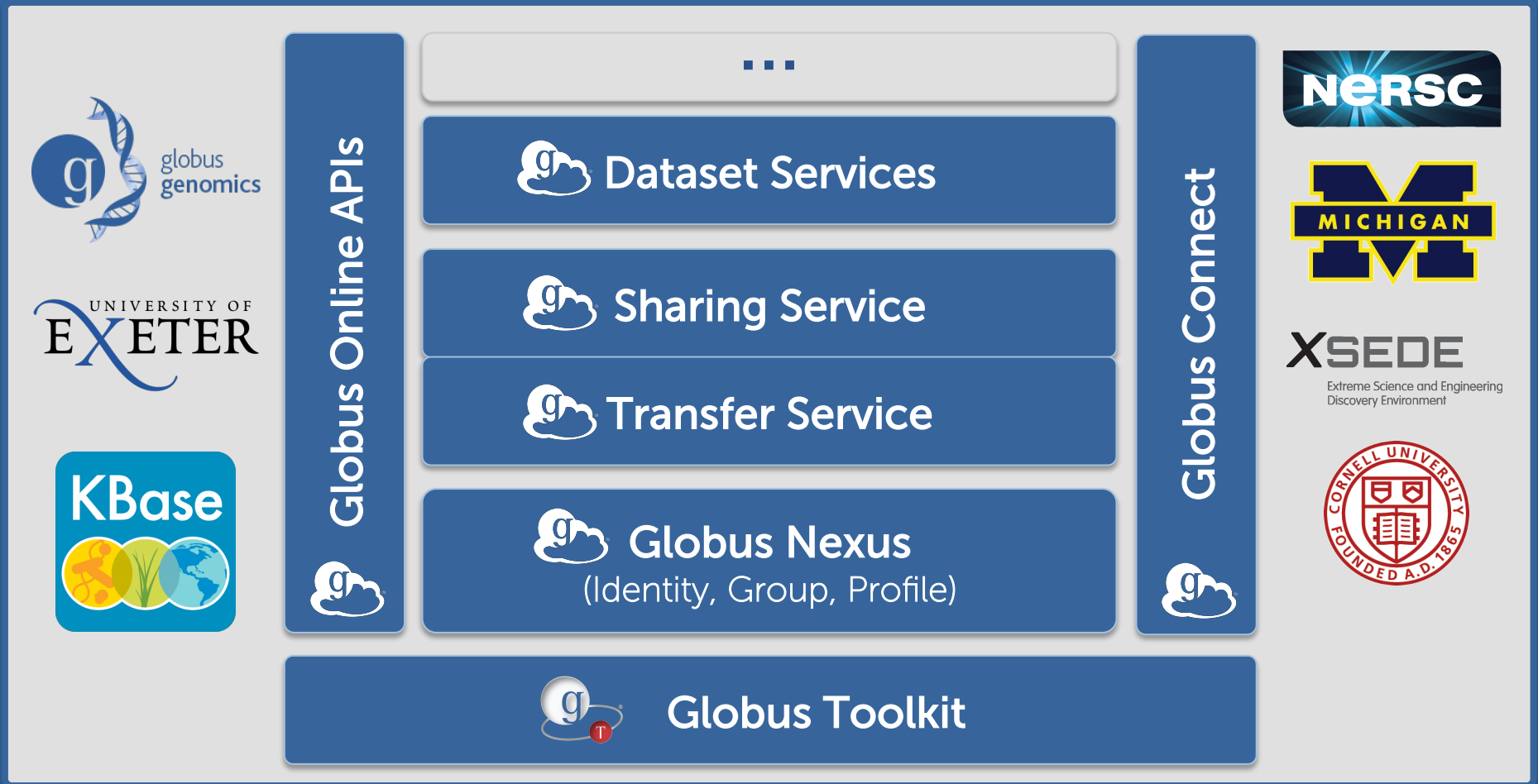
Starting at $20k per year

End User Plans

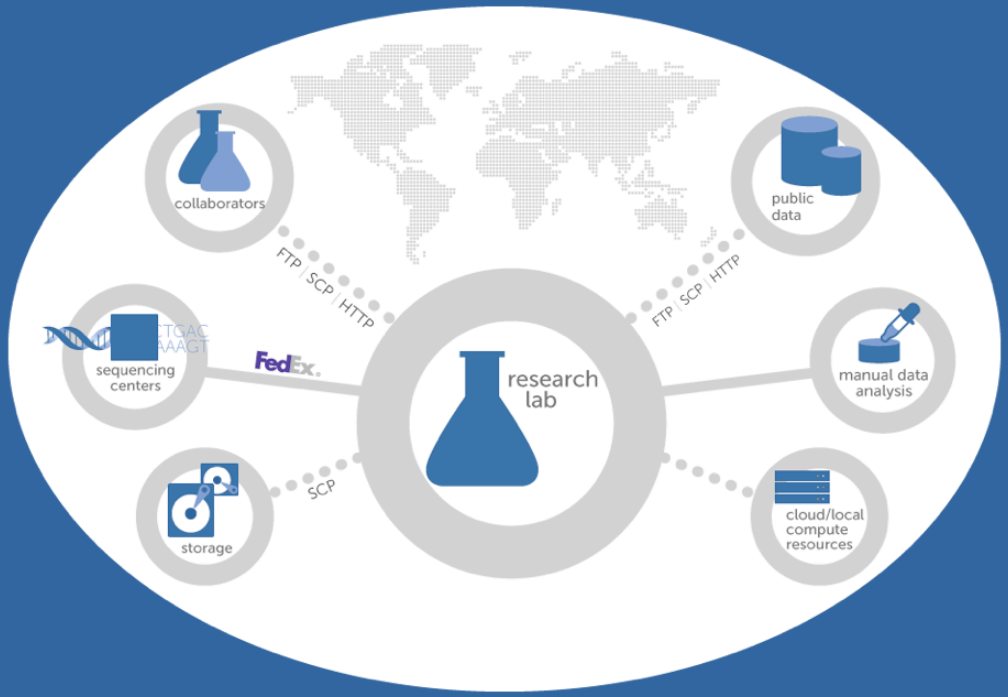
- **Basic: Free**
  - File transfer and synchronization to/from servers
  - Server endpoints with Globus Connect Multi-User
    - Can host shared endpoints for Plus subscribers
  - Personal endpoints with Globus Connect
  - Access to shared endpoints created by others

- **Plus: $7/month (or $70/year)**
  - Create and manage shared endpoints (from any sharable or personal endpoint)
  - Peer-to-peer (Globus Connect to Globus Connect)
  - Support for web and command line interfaces

Globus Platform-as-a-Service



Genomics research faces massive data management and analysis hurdles

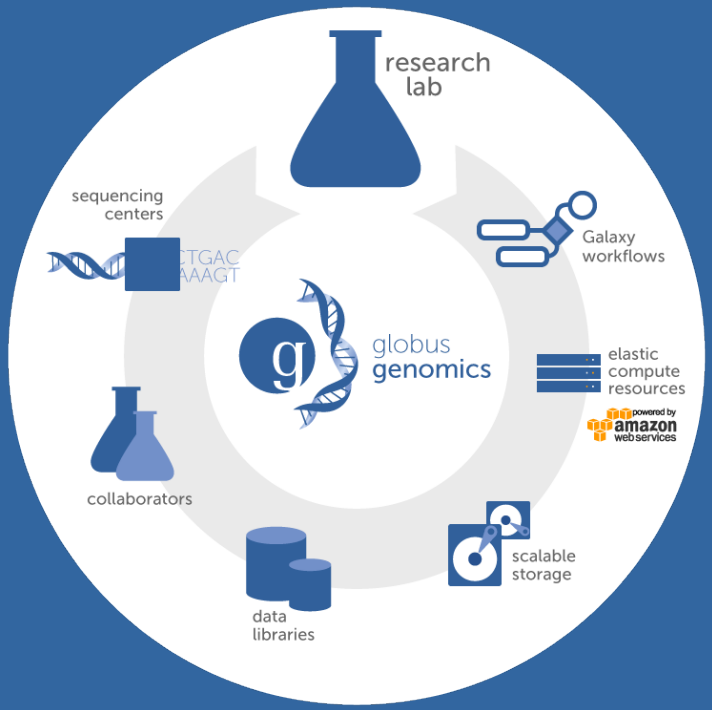


Globus Genomics

End-to-end sequencing anal

Flexible. Scalable. Simplified.

More details later today



globus.org/**genomics**

Our research is supported by:

U.S. DEPARTMENT OF
ENERGY

THE UNIVERSITY OF
CHICAGO

Argonne
NATIONAL LABORATORY

Thank you to our sponsors!

Gold Sponsor

EMC ISILON

Silver Sponsor

DataDirect™ NETWORKS

Program Preview

- **Wednesday**
  - Globus Online Experiences: UExeter, NERSC, UMichigan, TU Dortmund
  - Product Previews: Metadata, Genomics
  - Deep Dives: ESnet, NCSA, Fermilab, KBase
- **Thursday**
  - Keynote: David Lifka – Cornell University
  - Product Roadmap Update
  - Provider Spotlight: UChicago, PNNL
  - Community Updates: EGCF, OGF, SDSC, Indiana

Questions
Discussion

Globus Toolkit releases in past year

- **26 Apr 2012 GT 5.2.1**
  - Allow/deny paths a GridFTP server may access
  - GridFTP support for setting file modification time
- **24 Jul 2012 GT 5.2.2**
  - GridFTP hybrid independent/striped server
- **3 Dec 2012 GT 5.2.3**
  - GridFTP fixed logging bugs
  - GRAM support for LSF
- **13 Feb 2013 GT 5.2.4**
  - GridFTP sharing support beta
  - GridFTP make delegation optional

Globus Toolkit Future Plans

- **Globus Connect Multiuser as GT package**
- **GridFTP**
  - HDFS support (beta)
  - UDP/UDT w/ NAT traversal (alpha)
  - HTTP support (alpha)
  - Firewall friendly, single-port server (prototype)
  - Improved mass storage system support
- **GRAM5**
  - More schedulers (e.g., SLURM)
  - More scale and reliability
  - Prototype: JSDL over REST

Focus on research, not IT

- Eliminate data transfer, sharing, and management challenges
- Leverage best-practice analysis pipelines (RNA-Seq, Exome-Seq, ChIP-Seq, etc.)
- Develop custom pipelines with full control over algorithms, applications, and parameters
- Dramatically reduce sequencing analysis turnaround time
- Institutionalize bioinformatics expertise