

Globus for Research Data Management

Presented to
ATPESC 2017 Participants

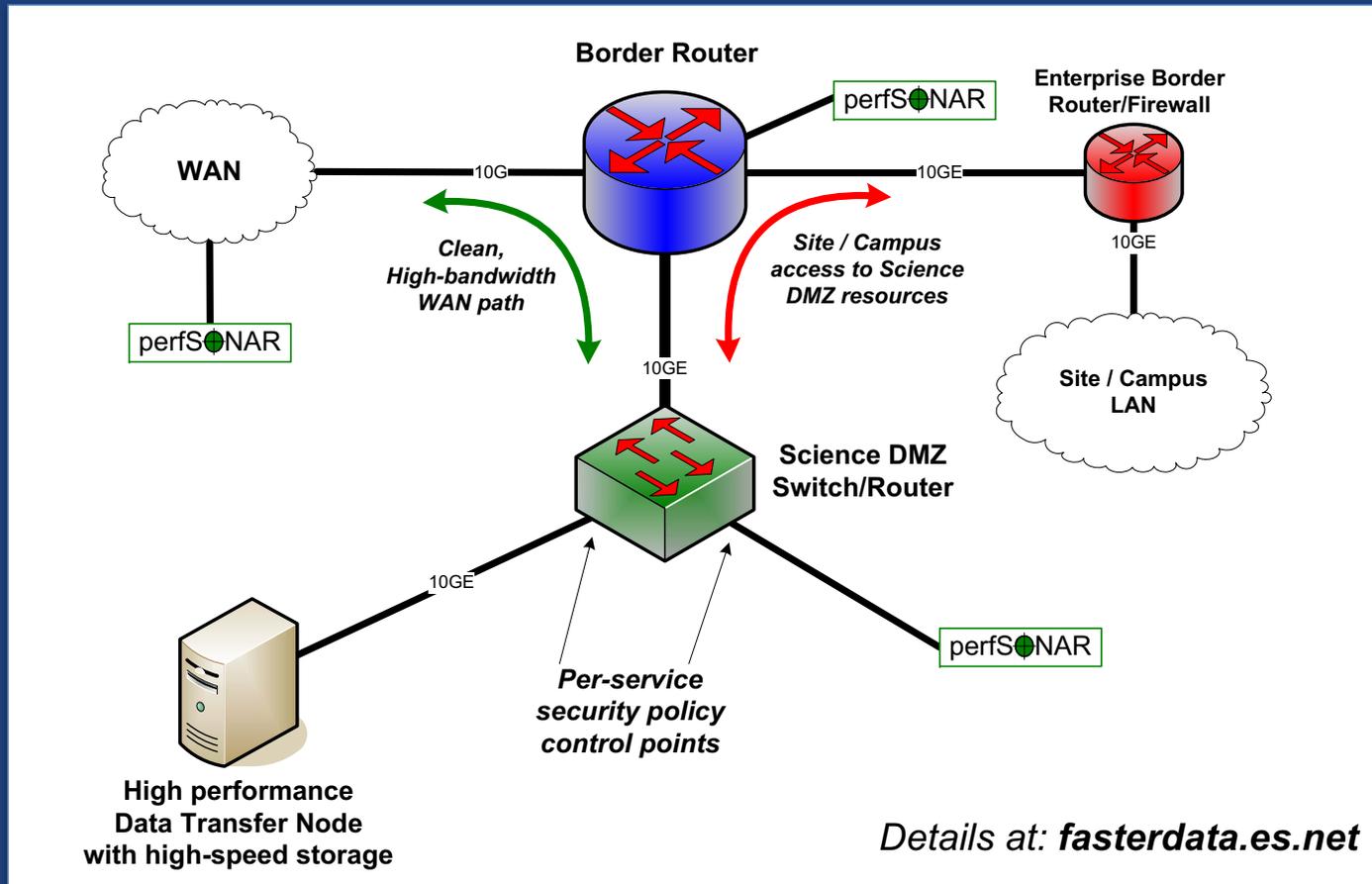
Greg Nawrocki
University of Chicago - Globus

Q Center, St. Charles, IL (USA)
Date 08/04/2017



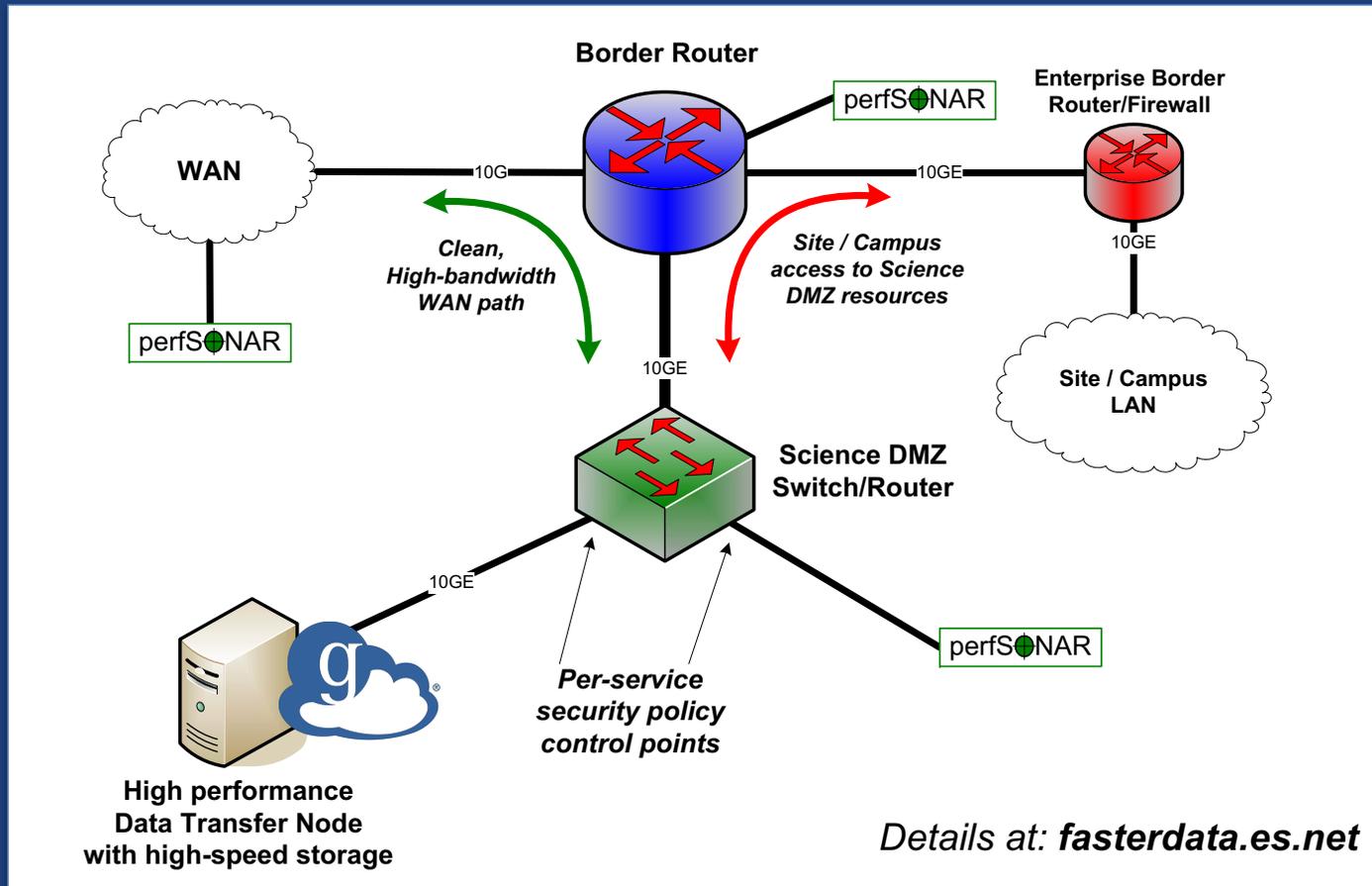


Best-practice deployment





Best-practice deployment





Research data management today



Index?



How do we...
...move?
...share?
...discover?
...reproduce?



Globus delivers...

Fast and reliable big data transfer,
sharing, publication, and discovery...

...directly from your own storage
systems...

...via software-as-a-service using existing
identities.



Globus enables...

Campus Bridging

...within and beyond campus
boundaries

Bridge to campus HPC

Move datasets to campus research computing center



Move results to laptop, department, lab...



Bridge to national cyberinfrastructure

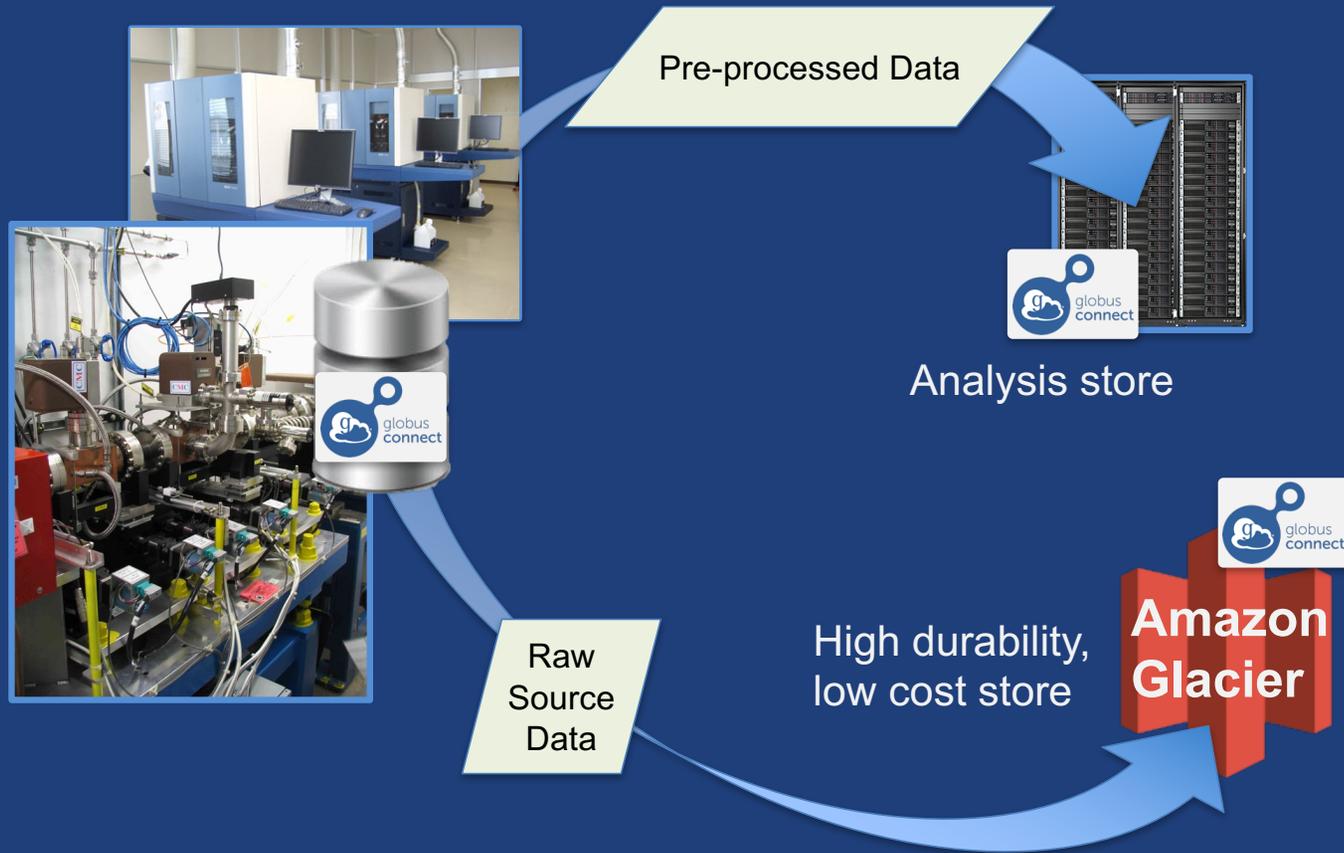
Move datasets to supercomputer,
national facility



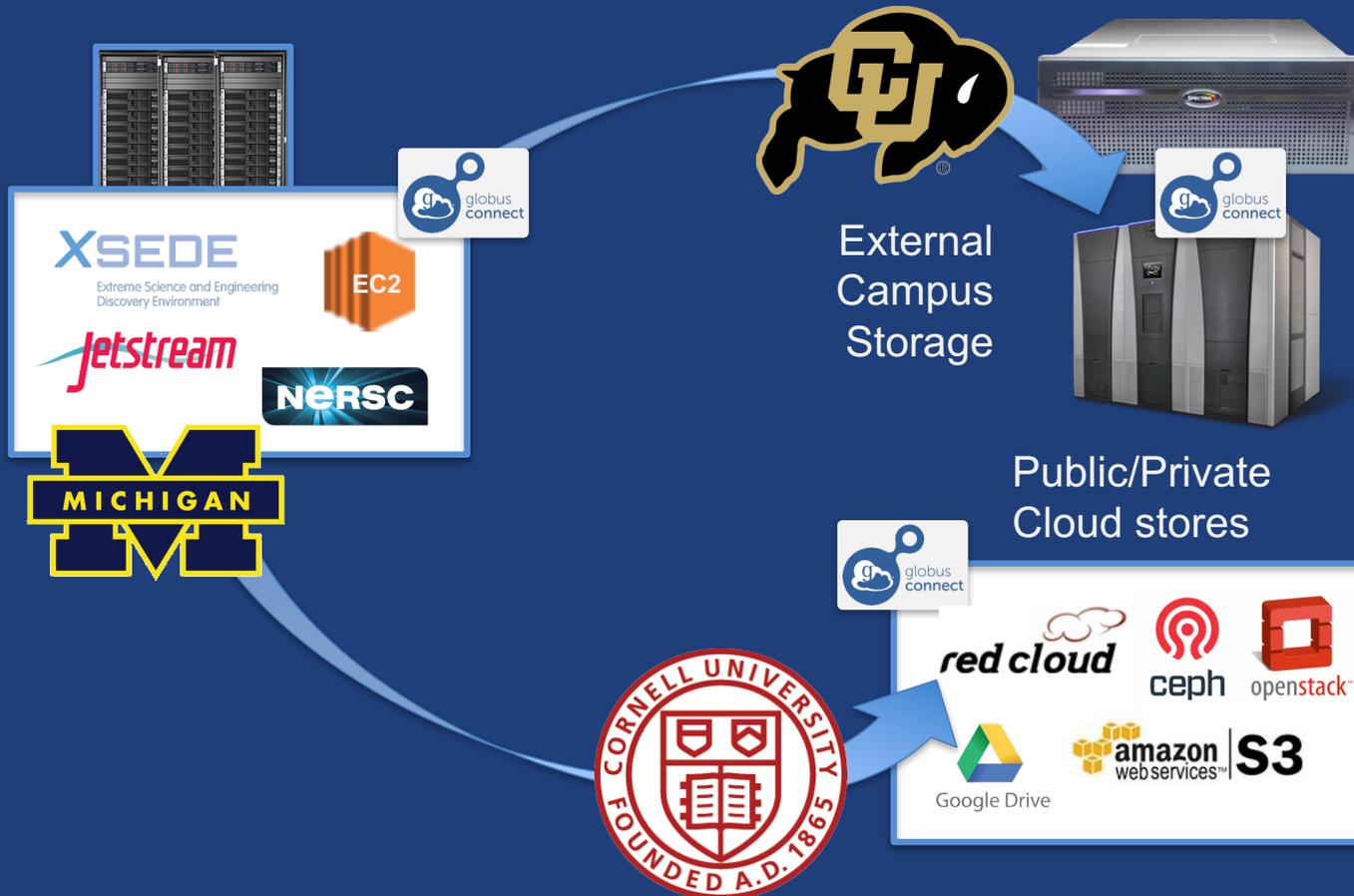
Move results to campus...



Bridge to instruments

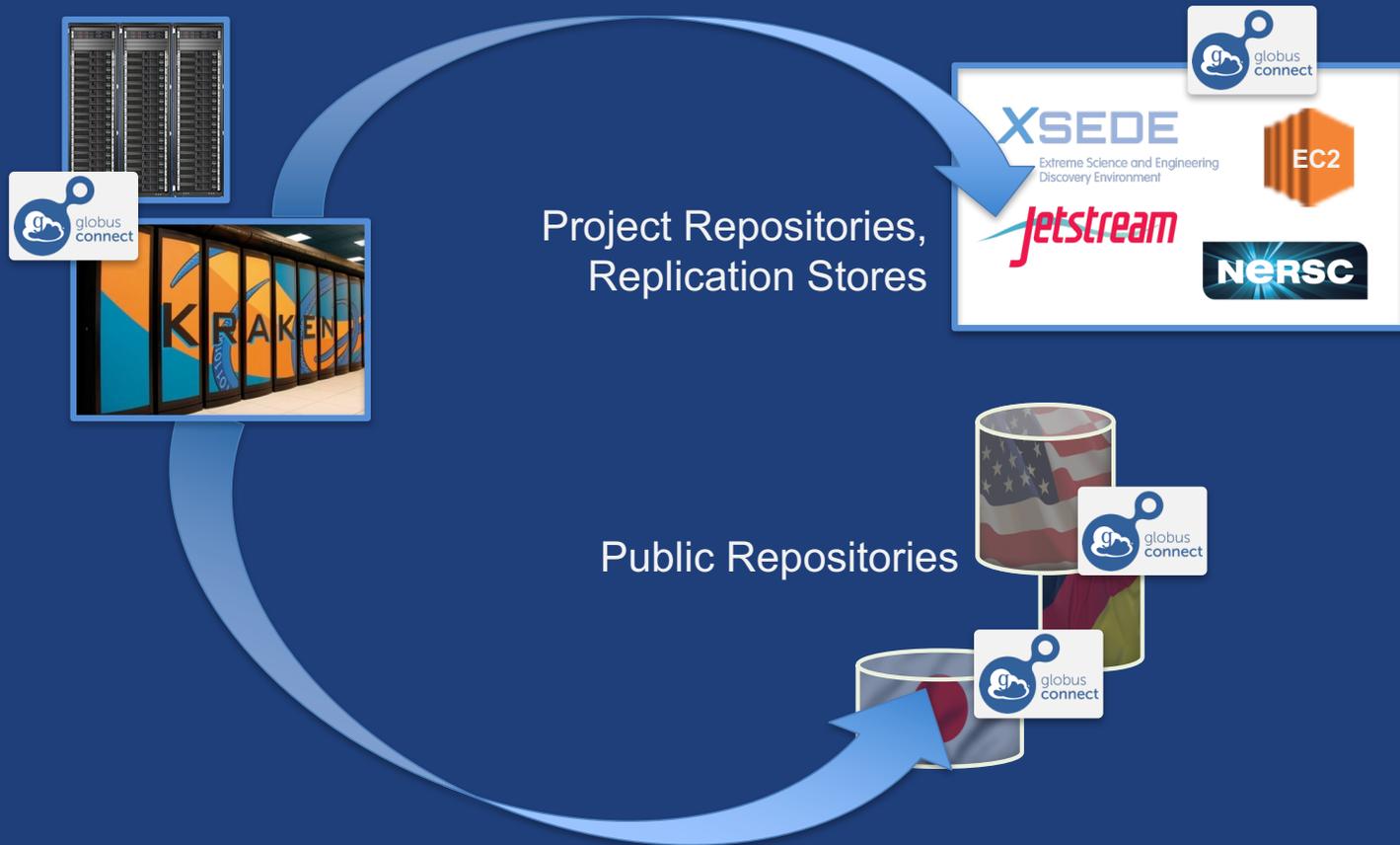


Bridge to collaborators





Bridge to community/public





Why use Globus?

- **Simplicity**
 - Consistent UI across systems
 - Easy access to collaborators
- **Reliability and performance**
 - “Fire-and-forget” file transfer
 - Maximized WAN throughput
- **Operational efficiency**
 - Low overhead SaaS model
 - Highly automatable: CLI, RESTful API
- **Access to a large and growing community**

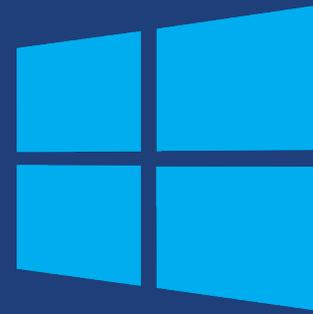


Endpoints

- **Storage abstraction**
 - All transfers happen between endpoints
 - Globus Connect Server instantiates endpoints
 - <https://docs.globus.org/faq/globus-connect-endpoints/>
- **Test / Demo Endpoints**
 - Globus Tutorial Endpoint 1
 - Globus Tutorial Endpoint 2
 - ESnet Test Endpoints
 - Read Only & Read / Write
 - Some contain file samples of various sizes
- **Globus Connect Personal**
 - Now your laptop is an endpoint
 - <https://www.globus.org/globus-connect-personal>



Globus Connect Personal (GCP)



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



Globus SaaS Demo Identities

- **Greg at Globus**
 - Globus ID: `nawrocki@globusid.org`
 - Email: `greg@globus.org`
- **Greg at University of Chicago**
 - CILogon: `nawrocki`
 - Email: `nawrocki@uchicago.edu`
- **Greg at home**
 - Globus ID: `nawrockipersonal@globusid.org`
 - Email: `greg@nawrockinet.com`



Globus SaaS Demo Identities

- **Greg at Globus** ← Primary Identity
 - Globus ID: nawrocki@globusid.org
 - Email: greg@globus.org
- **Greg at University of Chicago**
 - CILogon: nawrocki
 - Email: nawrocki@uchicago.edu

Linked Identities

- **Greg at home**
 - Globus ID: nawrockipersonal@globusid.org
 - Email: greg@nawrockinet.com



globus

Products ▾

Pricing

Developers

Support ▾

Log In



Research data management simplified.



TRANSFER



SHARE



PUBLISH



BUILD



Get unified access to your research data, across all systems, using any existing identity.

Laptop? HPC cluster? Cloud storage? Tape archive? Access them all using just a web browser.

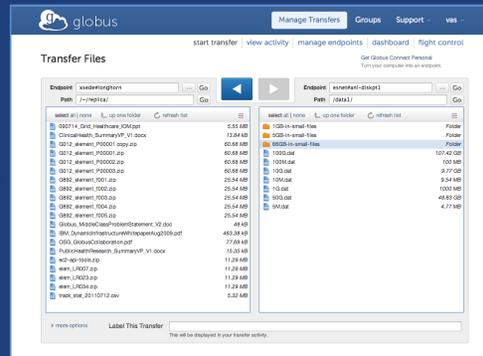
Data stored at a different institution? At a



Use(r)-appropriate interfaces



Globus service



Web

```
(globus-cli) jupiter:~ vas$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose          Control level of output
  -h, --help            Show this message and exit.
  -F, --format [json|text] Output format for stdout. Defaults to text
  --map-http-status TEXT Map HTTP statuses to any of these exit codes:
                        0,1,50-99. e.g. "404=50,403=51"

Commands:
  bookmark      Manage Endpoint Bookmarks
  config        Modify, view, and manage your Globus CLI config.
```

CLI

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

Rest API



How can I integrate
Globus into my
research workflows?



Globus serves as...

A platform for building science gateways, portals and other web applications in support of research and education.



Command Line Interface

- Transfer and Auth
- Replaces old SSH-based command line shell
- Uses Python SDK
- Open source

github.com/globus/globus-cli

docs.globus.org/cli

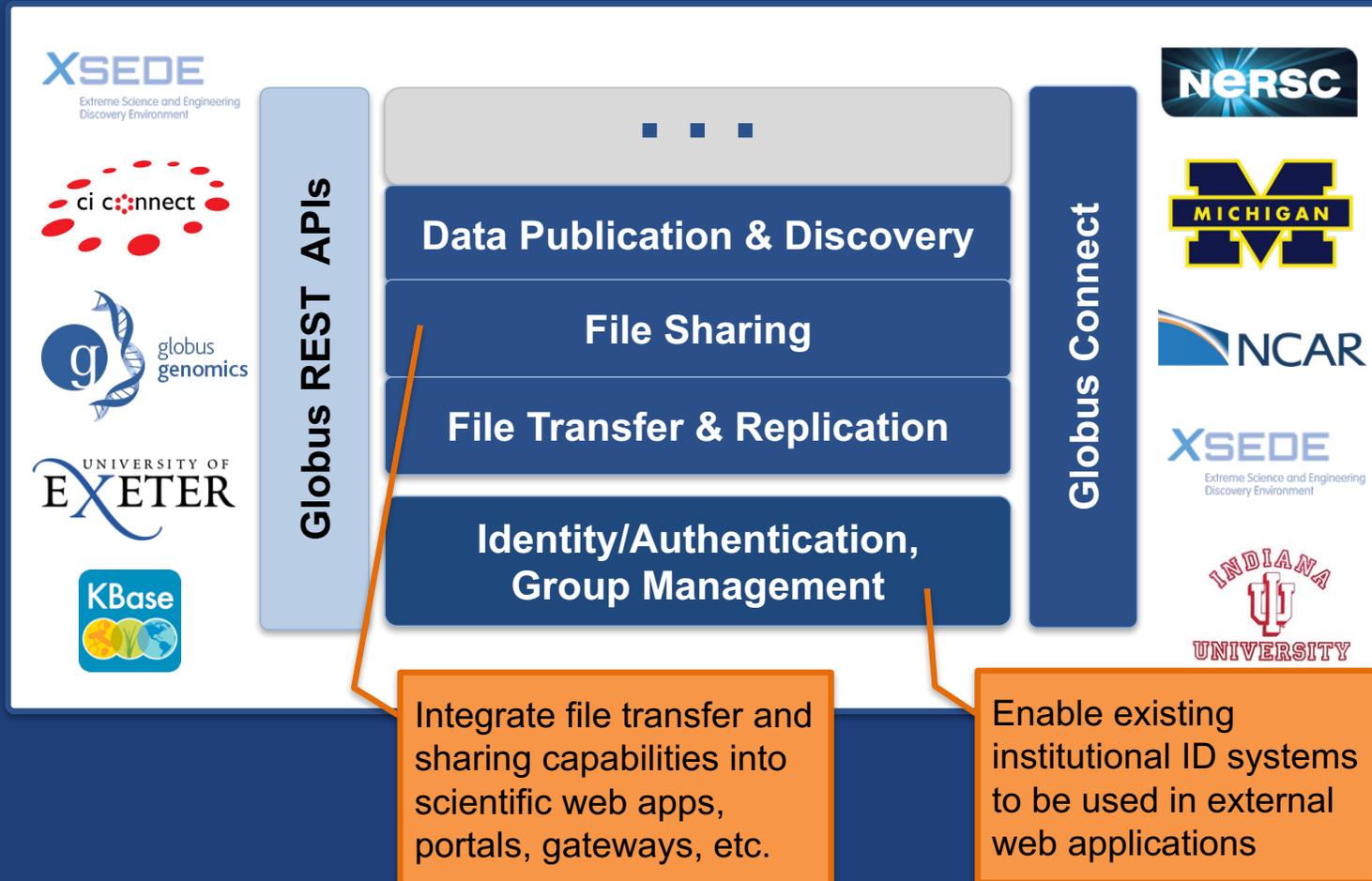
```
$ globus
Usage: globus [OPTIONS] COMMAND [ARGS]...

Options:
  -v, --verbose           Control level of output
  -h, --help             Show this message and exit.
  -F, --format [json|text] Output format for stdout. Defaults to text
  --jmespath, --jq TEXT  A JMESPath expression to apply to json output.
                        Takes precedence over any specified '--format' and
                        forces the format to be json processed by this
                        expression
  --map-http-status TEXT Map HTTP statuses to any of these exit codes:
                        0,1,50-99. e.g. "404=50,403=51"

Commands:
  bookmark      Manage Endpoint Bookmarks
  config        Modify, view, and manage your Globus CLI config.
  delete        Submit a Delete Task
  endpoint      Manage Globus Endpoint definitions
  get-identities Lookup Globus Auth Identities
  list-commands List all CLI Commands
  login         Login to Globus to get credentials for the Globus CLI
  logout        Logout of the Globus CLI
  ls            List Endpoint directory contents
  mkdir         Make a directory on an Endpoint
  rename        Rename a file or directory on an Endpoint
  task          Manage asynchronous Tasks
  transfer      Submit a Transfer Task
  version       Show the version and exit
  whoami        Show the currently logged-in identity.
```



Globus as PaaS





Data App: NCAR RDA

UCAR NCAR Closures/Emergencies Locations/Directions Find Pe

Hello [twecke@uchicago.edu](#) [dashboard](#) [sign out](#)

NCAR | Research Data Archive
UCAR | Computational & Information Systems Lab *weather • data • climate*

Go to Dataset:

[Home](#) [Find Data](#) [Ancillary Services](#) [About/Contact](#) [Data Citation](#) [Web Services](#) [For Staff](#)

NCEP Climate Forecast System Version 2 (CFSv2) Monthly Products

ds094.2

For assistance, contact [Bob Dattore](#) (303-497-1825).

[Description](#) [Data Access](#)

Mouse over the table headings for detailed descriptions

Data Description		Data File Downloads		Customizable Data Requests	Other Access Methods	NCAR-Only Access	
		Web Server Holdings	Globus Transfer Service (GridFTP)	Subsetting	THREDDS Data Server	Central File System (GLADE) Holdings	Tape Archive (HPSS) Holdings
Union of Available Products		Web File Listing	Request Globus Invitation	Get a Subset	TDS Access	GLADE File Listing	HPSS File Listing
P R O D	Diurnal monthly means	Web File Listing		Get a Subset		GLADE File Listing	HPSS File Listing
	Regular monthly means	Web File Listing		Get a Subset		GLADE File Listing	HPSS File Listing



Storage connectors

- **Standard storage connectors (Posix)**

- Linux, Windows, MacOS
- Lustre, GPFS, OrangeFS, etc.

- **Premium storage connectors**

AWS S3

Ceph RadosGW (S3 API)

Spectra Logic BlackPearl

Google Drive

HPSS

HDFS (in progress)

iRODS (in progress)

HGST Active Archive (in progress)

docs.globus.org/premium-storage-connectors



Globus Connect Server

- Runs on Linux
 - CentOS 5, 6, and 7
 - Debian 7 and 8
 - Fedora 23 and 24
 - Red Hat Enterprise Linux 5, 6, and 7
 - Scientific Linux 5, 6, and 7
 - SuSE Linux Enterprise Server 11sp3
 - Ubuntu 12.04 LTS, 14.04 LTS, 15.10, and 16.04 LTS

<https://docs.globus.org/globus-connect-server-installation-guide/>



Performance and Reliability

- **Multiple DTNs per endpoint**
- **Network Use Tuning**
 - Concurrency
 - Parallelism
- **Network Use Options**
 - Minimal
 - Normal
 - Aggressive
 - Custom

https://docs.globus.org/globus-connect-server-installation-guide/#setting_endpoint_network_use_options

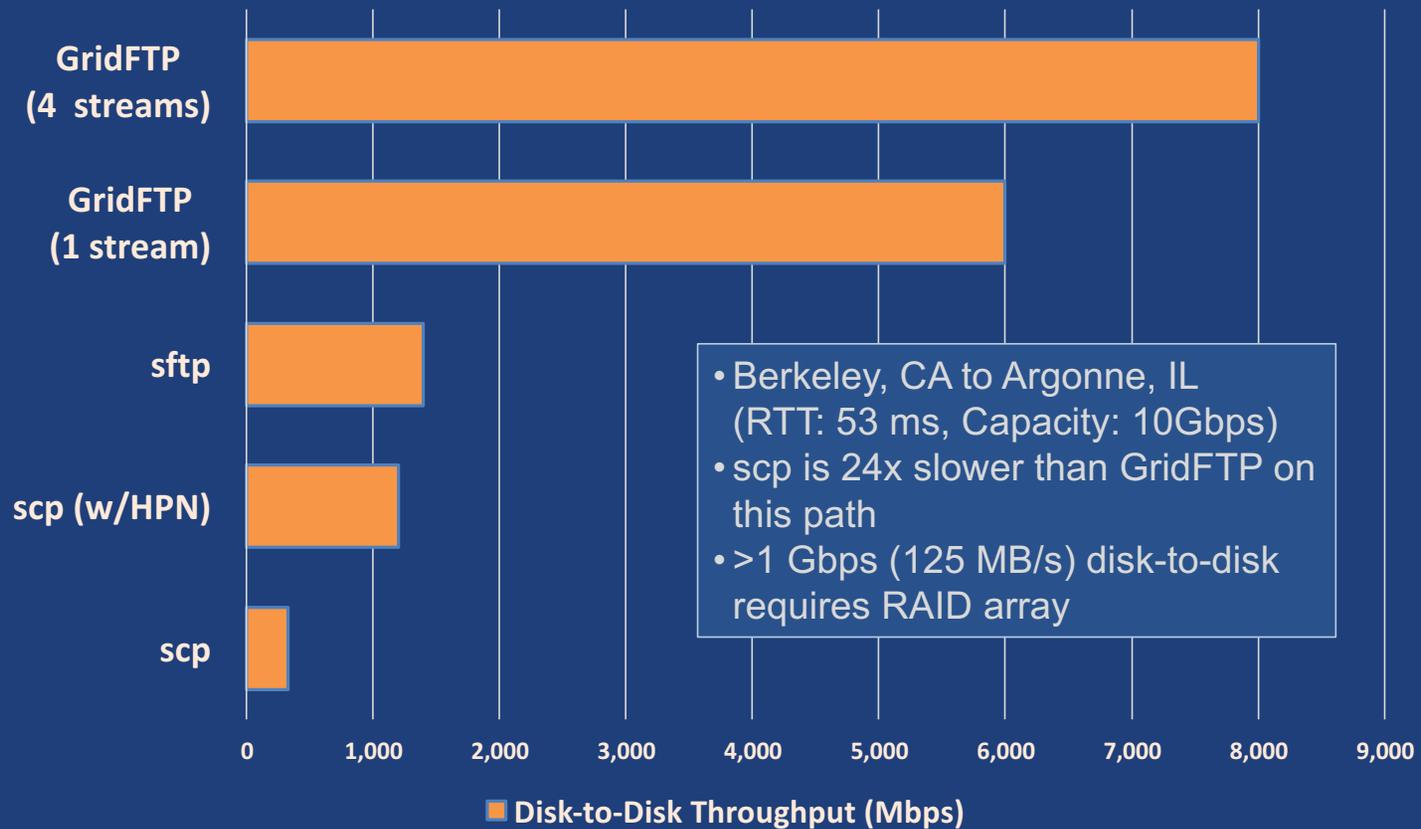


Illustrative performance

- **20x scp throughput (typical)**
 - >100x demonstrated
- **On par/faster than UDP based tools (NASA JPL study and anecdotal)**
- **Capable of saturating “any” WAN link**
 - Demonstrated 85Gbps sustained disk-to-disk
 - Typically require throttling for QoS



Disk-to-Disk Throughput



Source: ESnet (2016)



Thank you to our sponsors



U.S. DEPARTMENT OF
ENERGY

NIST

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



THE UNIVERSITY OF
CHICAGO



Argonne
NATIONAL LABORATORY



powered by
amazon
web services



Globus sustainability model

- **Standard Subscription**

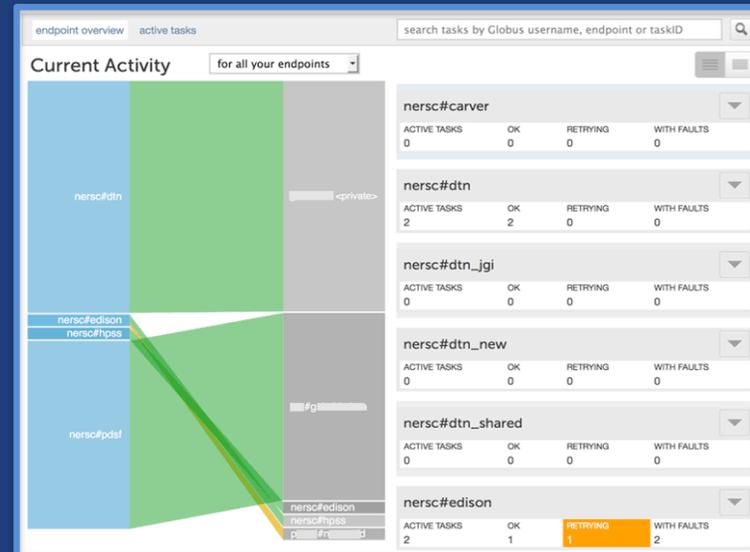
- Shared endpoints
- Data publication
- HTTPS support*
- Management console
- Usage reporting
- Priority support
- Application integration

- **Branded Web Site**

- **Premium Storage Connectors**

- Amazon S3, Ceph, HPSS, Spectra, Google Drive, Box*, HDFS*

- **Alternate Identity Provider (InCommon is standard)**



*Coming soon



Globus by the numbers

48

most server endpoints on one campus

280 PB
transferred

47 billion
tasks processed

60,000
registered users

350

100TB+ users

10,000
active users

3 months

longest running managed transfer

10,000

active endpoints

300+

federated identities

1 PB

largest single transfer to date

5,119

active shared endpoints

99.5%

uptime

THANK YOU, subscribers!





Join the Globus Community

- **Documentation**
 - docs.globus.org
- **Join the mailing lists**
 - globus.org/mailling-lists
- **Lots of good open source examples**
 - github.com/globus/
 - github.com/globus/globus-sdk-python
 - Discussions on developer-discuss@globus.org
- **When all else fails**
 - <https://www.globus.org/contact-us>



Globus Admin

- **Globus Connect Server (GCS) Installation**
 - <https://docs.globus.org/globus-connect-server-installation-guide/>
- **Globus Connect Server Installation on the EC2 Tutorial Server**
 - <https://www.globusworld.org/tutorials>
 - You'll need your own EC2 instance
 - When we do the tour we supply temporary instances
- **Helpful slides**
 - https://www.globusworld.org/files/2017/170124_GWTour_Globus_Admin_Tutorial.pdf
- **Configuration options**
 - /etc/globus-connect-server.conf



Automation Examples

- Syncing a directory
 - Bash script that calls the Globus CLI and a Python module that can be run as a script or imported as a module.
- Staging data in a shared directory
 - Bash / Python
- Removing directories after files are transferred
 - Python script
- Simple code examples for various use cases using Globus
 - <https://github.com/globus/automation-examples>



Globus Transfer API Set

- **Helpful slides**
 - https://www.globusworld.org/files/2017/170412_GW17_Dev_Tutorial.pdf
 - Both transfer and auth covered
- **Doc**
 - <https://docs.globus.org/api/transfer/>
- **Sample data portal**
 - <https://github.com/globus/globus-sample-data-portal>
- **Jupyter notebook**
 - <https://github.com/globus/globus-jupyter-notebooks>



Globus Auth API Set

- **Helpful slides**
 - https://www.globusworld.org/files/2017/170412_GW17_Dev_Tutorial.pdf
 - Both transfer and auth covered
- **Doc**
 - <https://docs.globus.org/api/auth/>
- **Sample data portal**
 - <https://github.com/globus/globus-sample-data-portal>
- **Native app examples**
 - <https://github.com/globus/native-app-examples>



Globus on your Campus

- **Webinars**
- **Programs**
 - Helping you evangelize Globus within your institution.
- **Professional Services**
- **Globus World Tour**
 - Taking the show on the road.



Globus for Research Data Management

ATPESC 2017

August 4, 2017

Greg Nawrocki
greg@globus.org

