Broadening Access to Cyberinfrastructure with Globus for Research Data Management and the Globus Platform

RMACC 2017

August 17, 2017

Greg Nawrocki
greg@globus.org

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

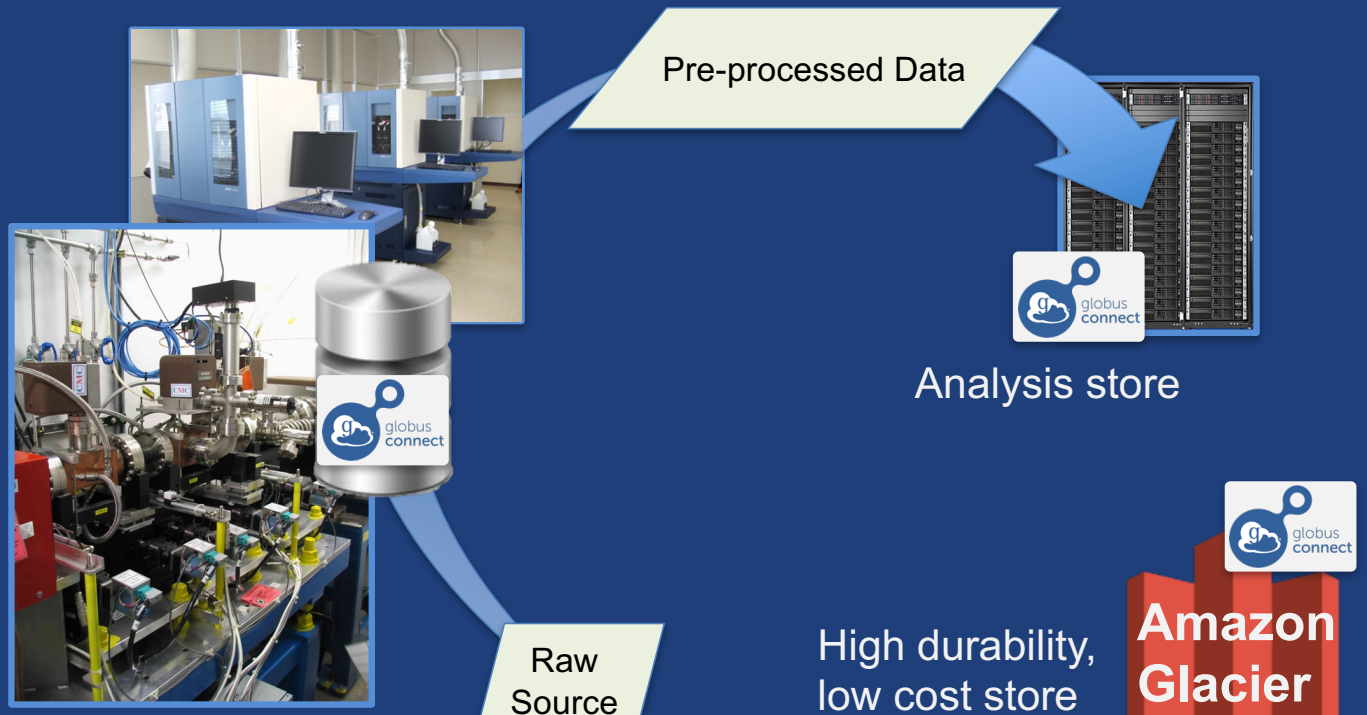
Move results to laptop, department, lab…

Bridge to national cyberinfrastructure

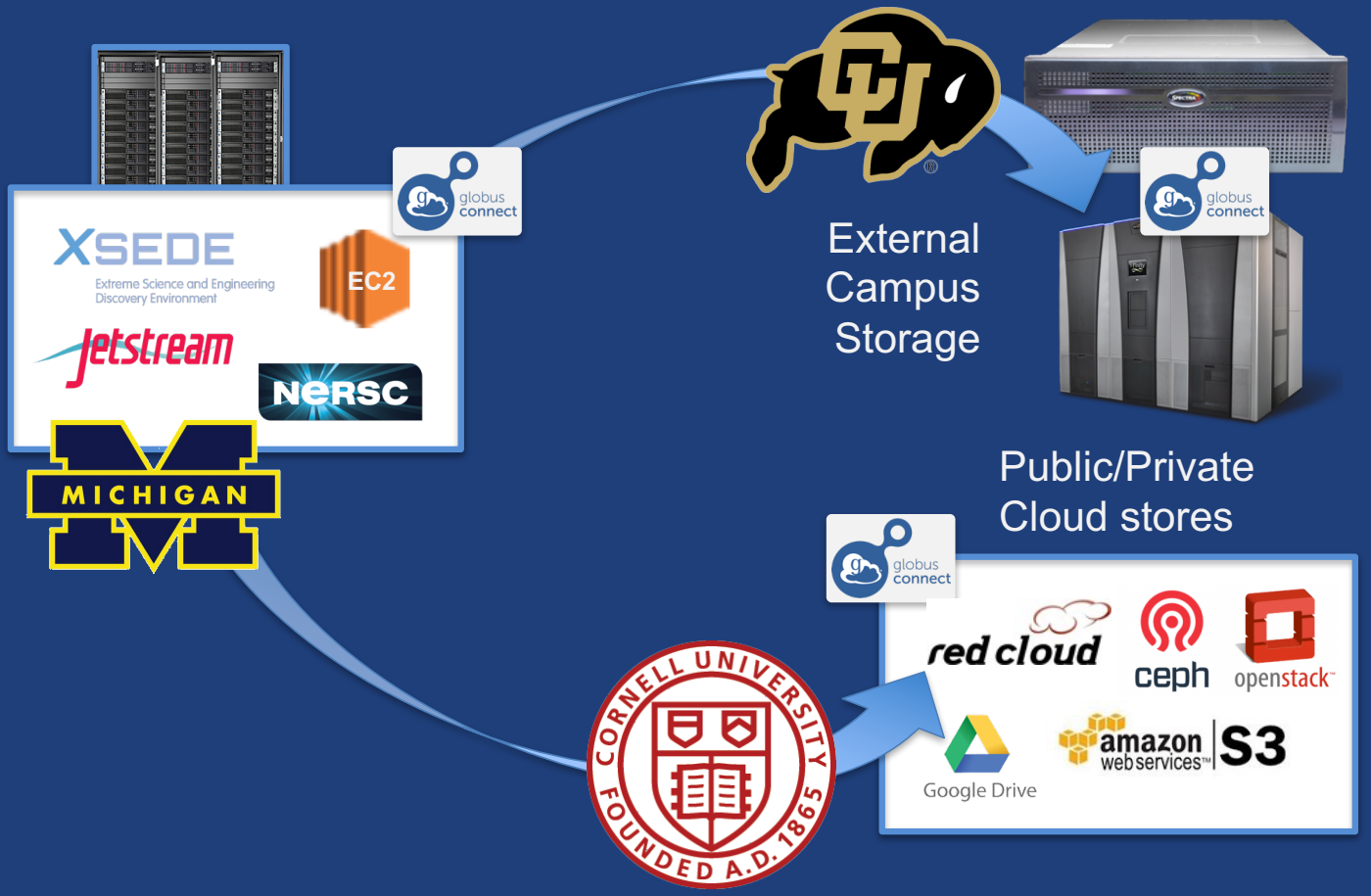
Move datasets to supercomputer, national facility

Move results to campus…

Bridge to instruments

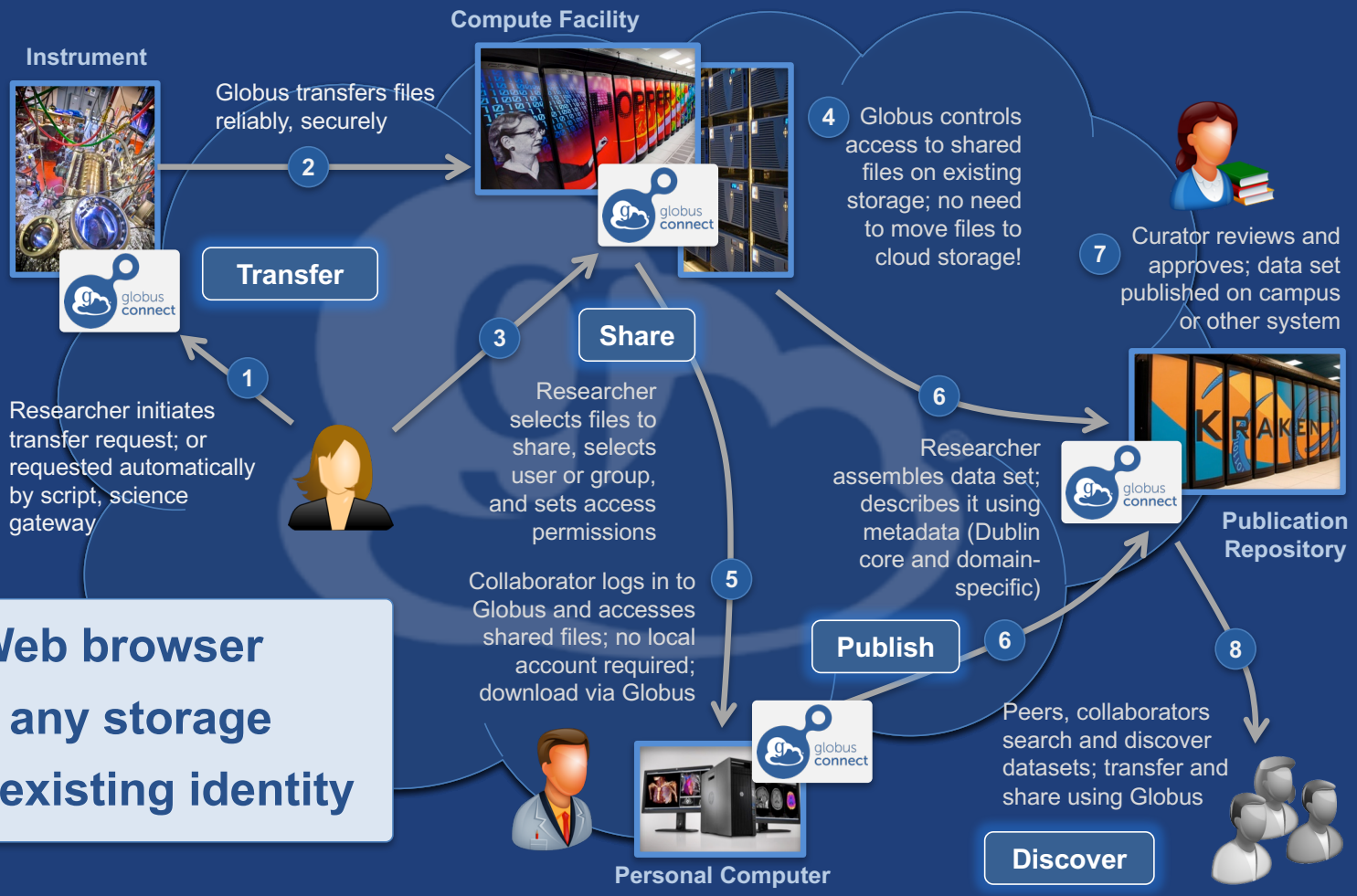


Bridge to collaborators



Bridge to community/public

Globus SaaS: Research data lifecycle



- Use a Web browser
- Access any storage
- Use an existing identity

Why use Globus?

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

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/

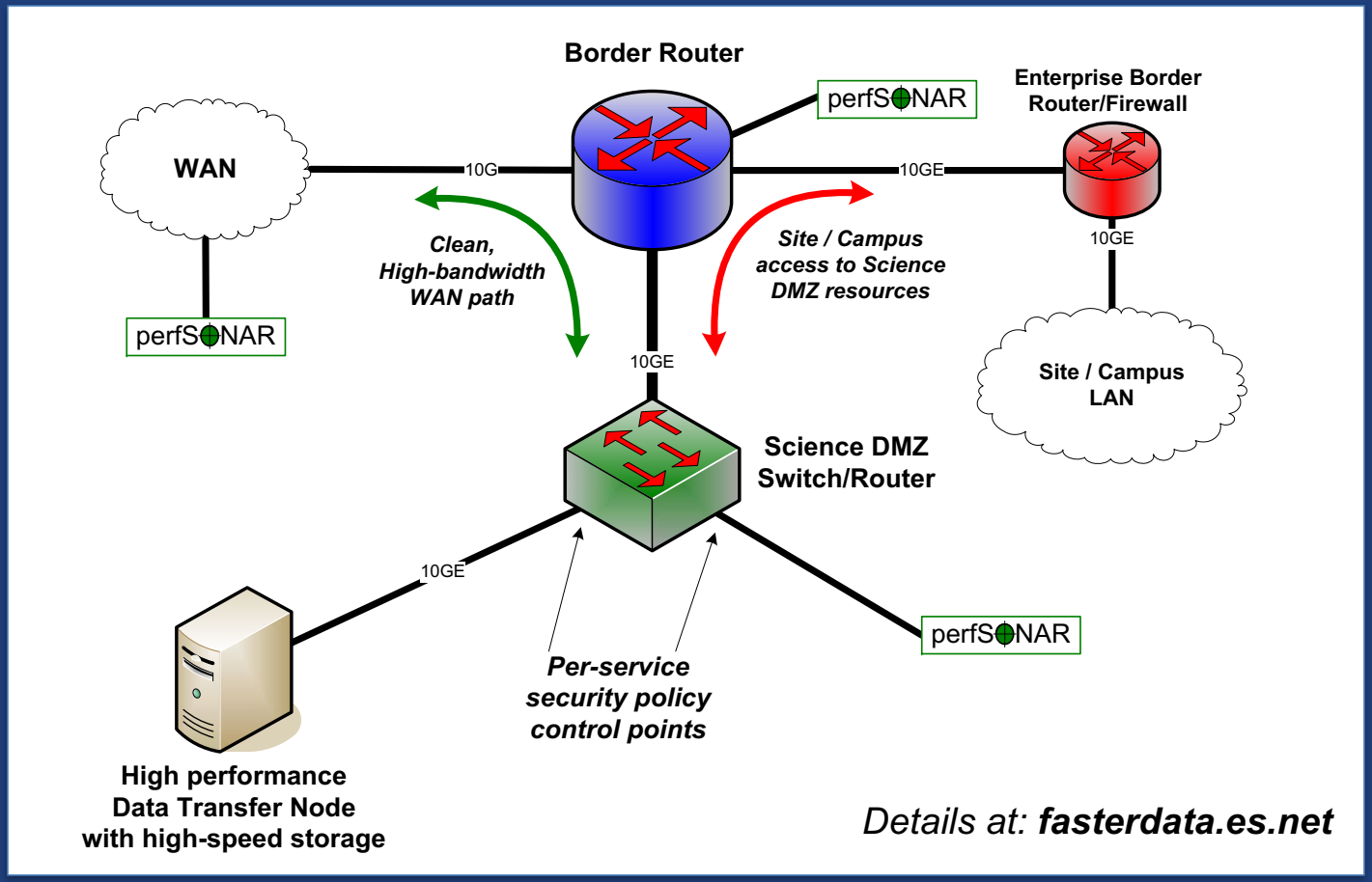
Storage connectors

- **Standard storage connectors (Posix)**
  - Linux, Windows, MacOS
  - Lustre, GPFS, OrangeFS, etc.
- **Premium storage connectors**

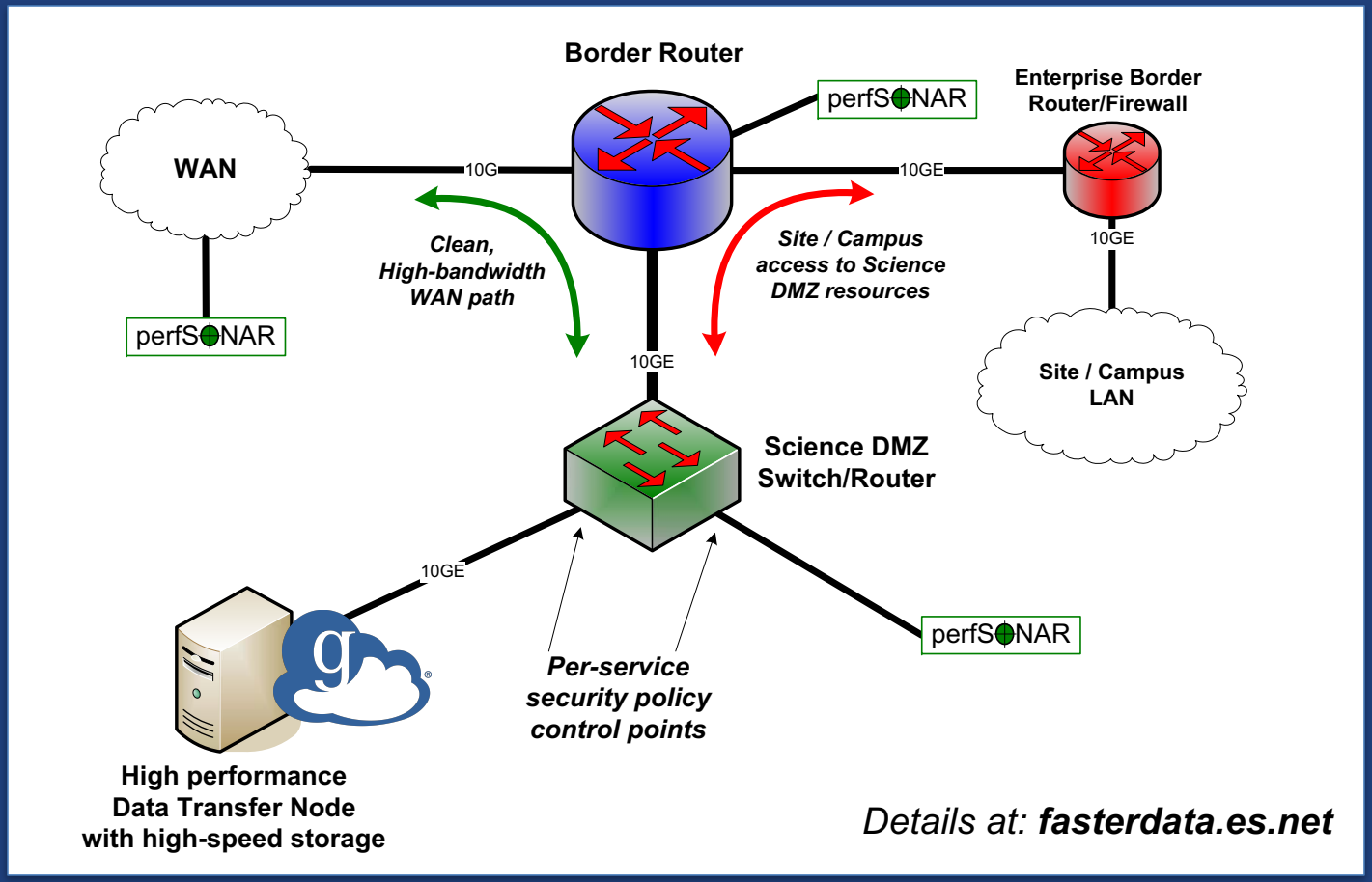
AWS S3	HPSS
Ceph RadosGW (S3 API)	HDFS (in progress)
Spectra Logic BlackPearl	iRODS (in progress)
Google Drive	HGST Active Archive (in progress)

docs.globus.org/premium-storage-connectors

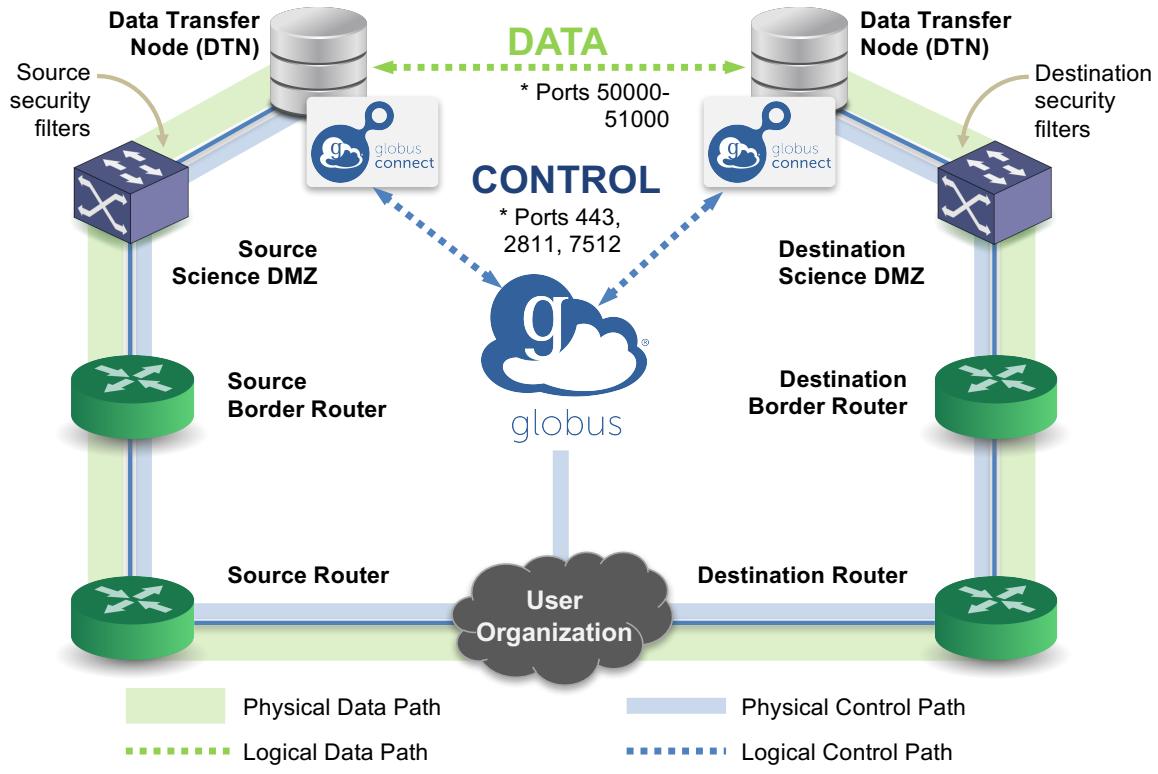
Best-practice deployment



Best-practice deployment



Network Paths - Illustrative



* Please see TCP ports reference: https://docs.globus.org/resource-provider-guide/#open-tcp-ports_section

Performance and Reliability

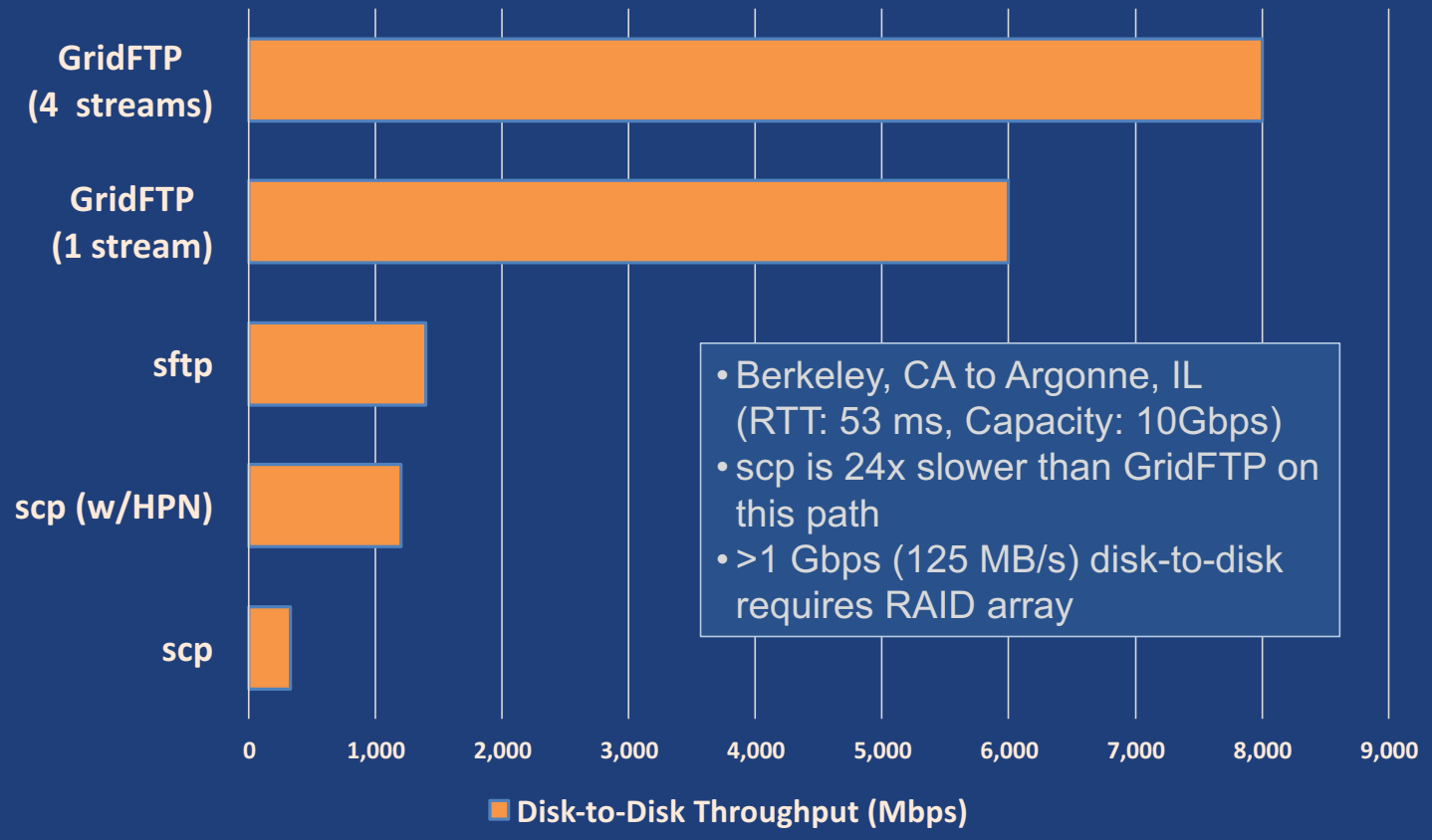
- **Multiple DTNs per endpoint**
- **Network Use Tuning**
  - Concurrency
  - Parallelism
- **Network Use Options**
  - Minimal
  - Normal
  - Aggressive
  - Custom

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)

docs.globus.org

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

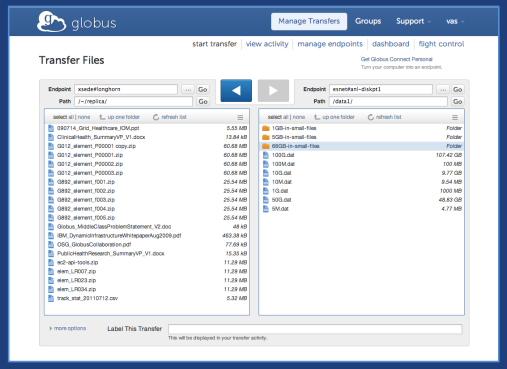
Linked Identities

- **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
high performance computing
local storage
institutional storage
commercial
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.
```

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

Cloud has transformed how software and platforms are delivered

Software as a service: **SaaS**

(web & mobile apps)

NETFLIX

Platform as a service: **PaaS**

Infrastructure as a service: **IaaS**

PaaS enables more rapid, cheap, and scalable delivery of powerful (SaaS) apps

Cloud has transformed how software and platforms are delivered

Software as a service: **SaaS**

(web & mobile apps)

Tripit from Concur

NETFLIX

salesforce

Platform as a service: **PaaS**

amazon web services

Microsoft Azure

force.com cloud platform

Infrastructure as a service: **IaaS**

amazon web services EC2

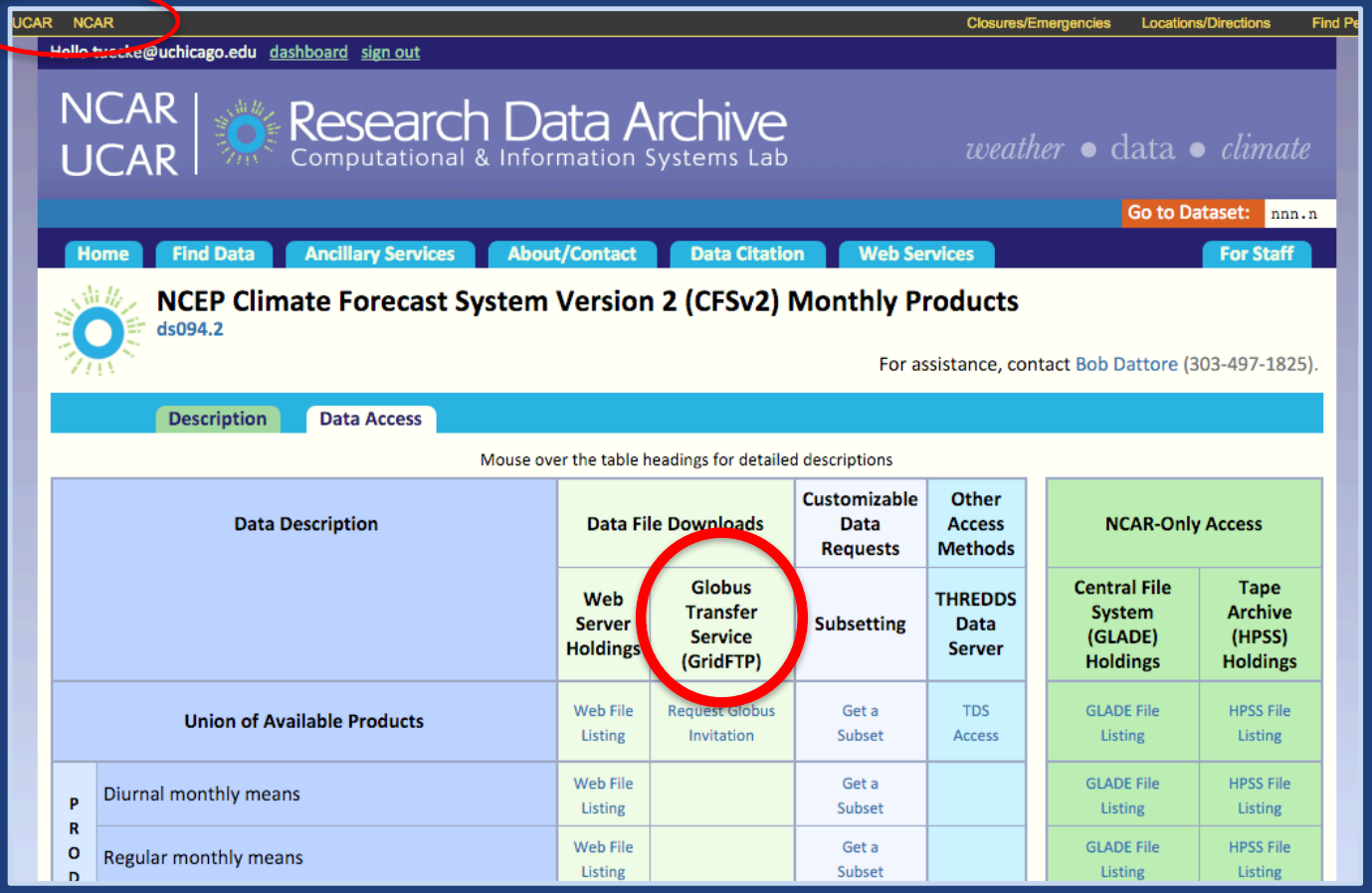
amazon web services S3

Microsoft Azure

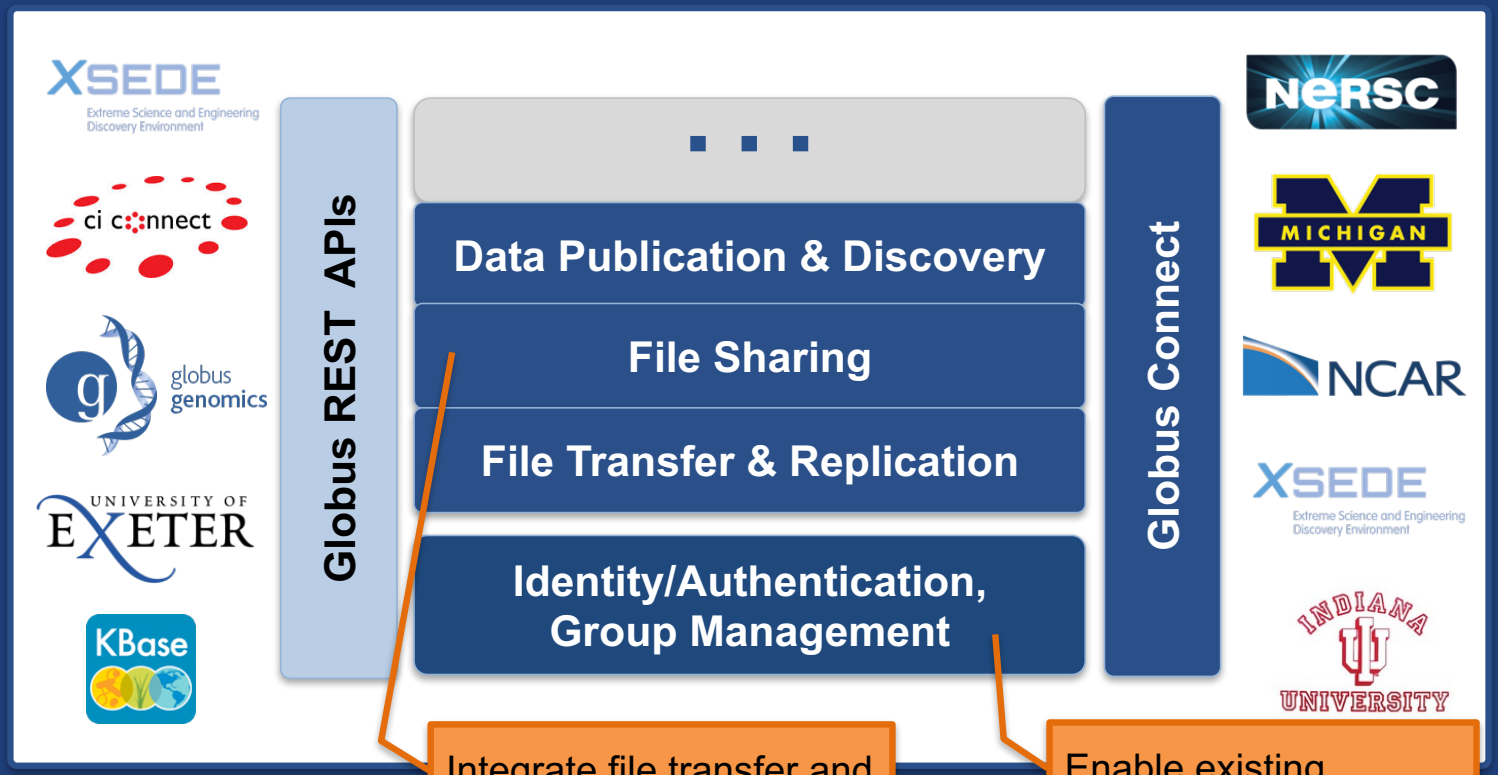
Google Compute Engine

PaaS enables more rapid, cheap, and scalable delivery of powerful (SaaS) apps

Data App: NCAR RDA



Globus as PaaS



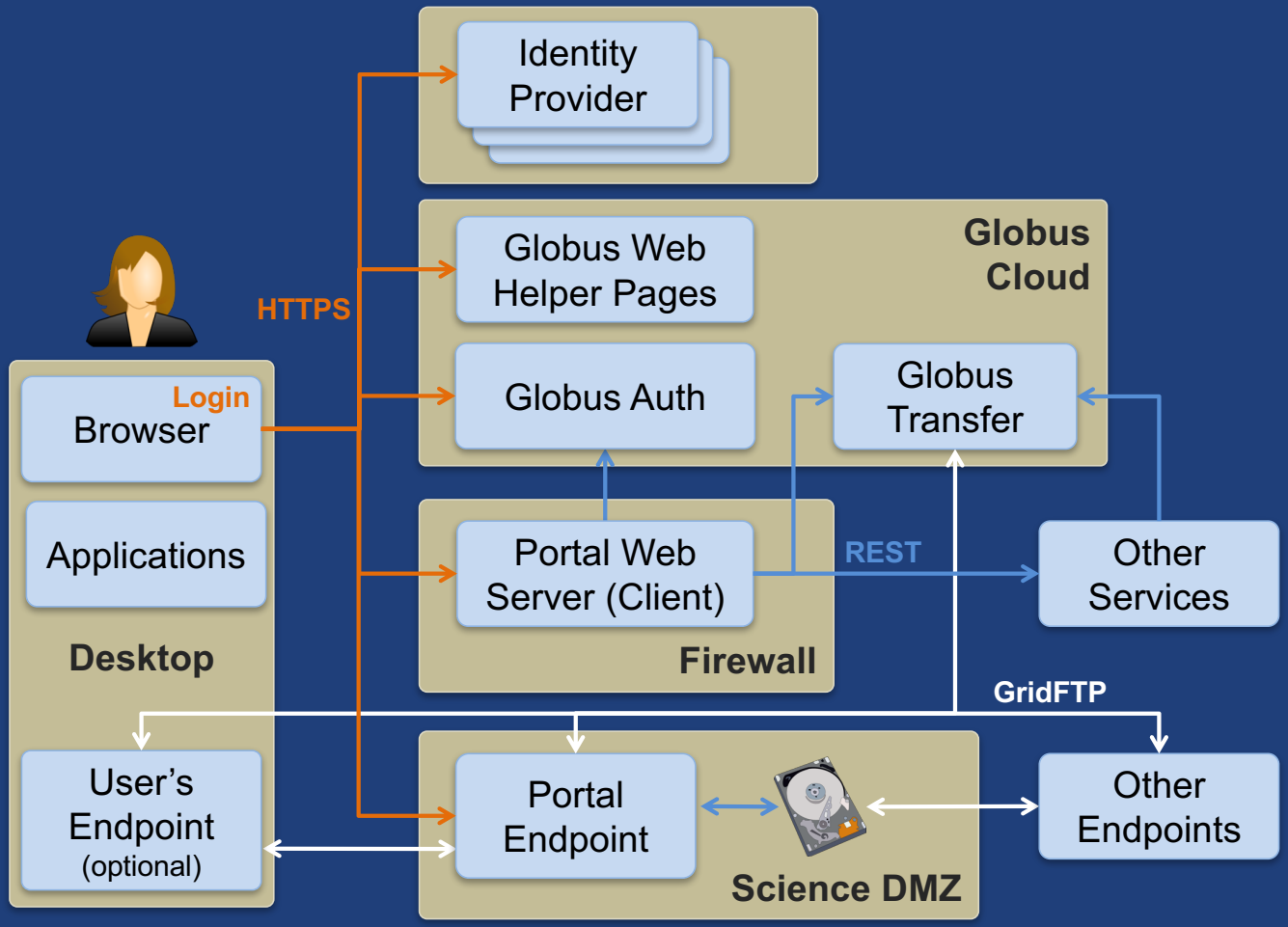
docs.globus.org

Demo

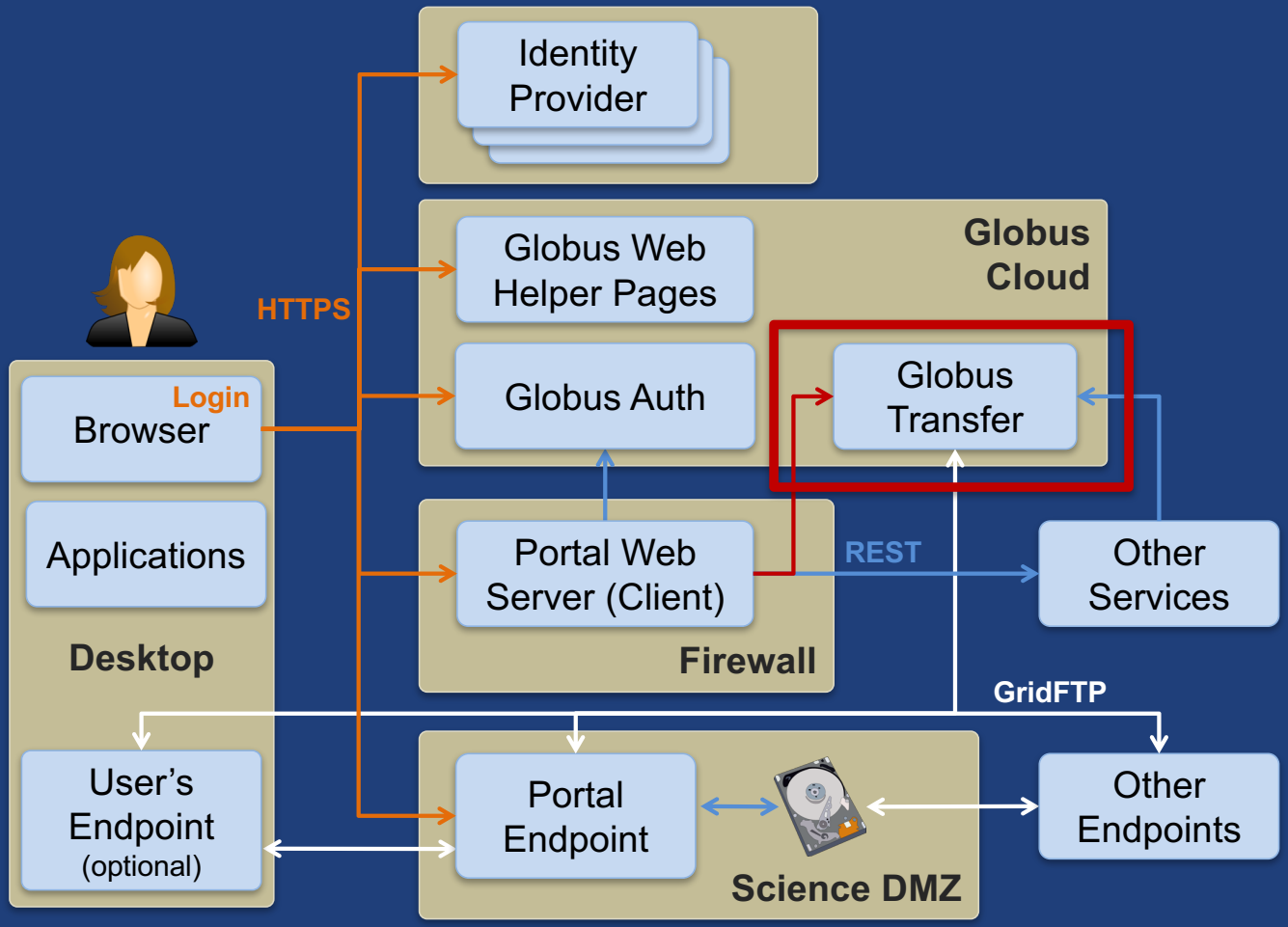
Sample
Research Data Portal

github.com/globus/globus-sample-data-portal

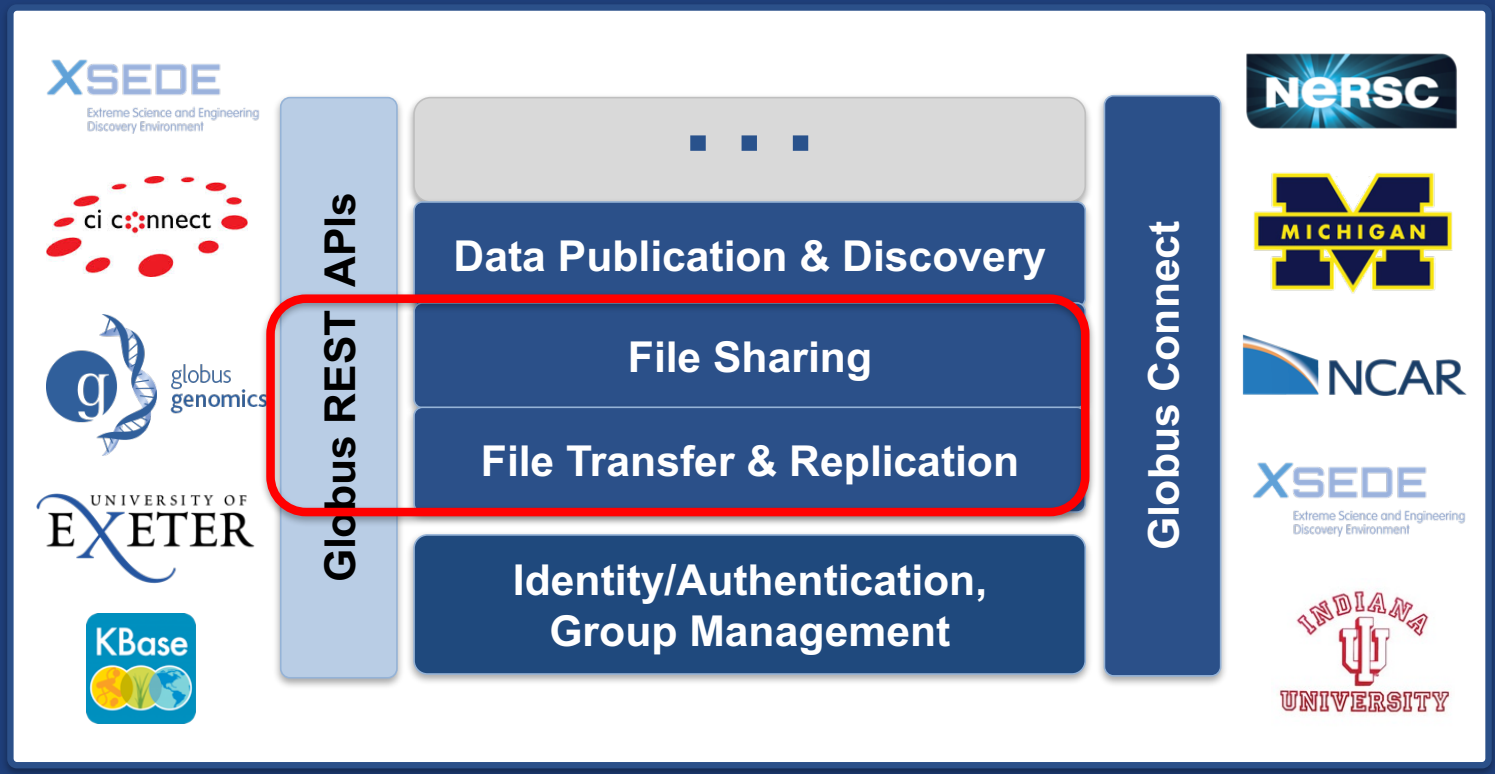
Prototypical research data portal



Prototypical research data portal



Globus as PaaS



Introduction to REST APIs

- **Remote operations on resources via HTTPS**
  - POST ~= Create (or other operations)
  - GET ~= Read
  - PUT ~= Update
  - DELETE ~= Delete
- **Globus APIs use JSON for documents and resource representations**
- **Resource named by URL**
  - Query params allow refinement (e.g., subset of fields)
- **Requests authorized via OAuth2 access token**
  - Authorization: Bearer asdflkqhafsdafeawk

Globus Transfer API

- **Nearly all Globus Web App functionality implemented via public Transfer API**

docs.globus.org/api/transfer

- **Stable, but small changes from time to time.**
  - Deprecation policy
  - developer-discuss@globus.org

Globus Python SDK

- **Python client library for the Globus Auth and Transfer REST APIs**

globus.github.io/globus-sdk-python

- **Public beta, likely to change some**

Python SDK Jupyter notebook

- **Jupyter (iPython) notebook demonstrating use of Python SDK**

github.com/globus/globus-jupyter-notebooks

- **Overview**
- **Open source, enjoy**

`TransferClient` class

- `globus_sdk.TransferClient` **class**

```
from globus_sdk import TransferClient
tc = TransferClient()
```

- **Handles connection management, security, framing, marshaling**

`TransferClient` low-level calls

- **Thin wrapper around REST API**
  – post(), get(), update(), delete()

```
get(path, params=None, headers=None, auth=None,
response_class=None)
```

  - `path` – path for the request, with or without leading slash
  - `params` – dict to be encoded as a query string
  - `headers` – dict of HTTP headers to add to the request
  - `response_class` – class for response object, overrides the client's default_response_class
  - Returns: `GlobusHTTPResponse` object

`TransferClient` higher-level calls

- **One method for each API resource and HTTP verb**
- **Largely direct mapping to REST API**

```
endpoint_search(filter_fulltext=None,
                filter_scope=None,
                num_results=25,
                **params)
```

Endpoint Search

- **Plain text search for endpoint**
  - Searches owner, display name, keywords, description, organization, department
  - Full word and prefix match
- **Limit search to pre-defined scopes**
  - `all, my-endpoints, recently-used, in-use, shared-by-me, shared-with-me`
- **Returns: List of endpoint documents**

Endpoint Management

- **Get endpoint (by id)**
- **Update endpoint**
- **Create & delete (shared) endpoints**
- **Manage endpoint servers**

Endpoint Activation

- **Activating endpoint means binding a credential to an endpoint for login**
- **Globus Connect Server endpoint that have MyProxy or MyProxy OAuth identity provider require login via web**
- **Auto-activate**
  - Globus Connect Personal and shared endpoints use Globus-provided credential
  - An endpoint that shares an identity provider with another activated endpoint will use credential
- **Must auto-activate before any API calls to endpoints**

File operations

- **List directory contents (ls)**
- **Make directory (mkdir)**
- **Rename**
- **Note:**
  - Path encoding & UTF gotchas
  - Don’t forget to auto-activate first

Task submission

- **Asynchronous operations**
  - Transfer
    - Sync level option
  - Delete
- **Get `submission_id`, followed by submit**
  - Once and only once submission

Task management

- **Get task by id**
- **Get task_list**
- **Update task by id (label, deadline)**
- **Cancel task by id**
- **Get event list for task**
- **Get task pause info**

Bookmarks

- **Get list of bookmarks**
- **Create bookmark**
- **Get bookmark by id**
- **Update bookmark**
- **Delete bookmark by id**

- **Cannot perform other operations directly on bookmarks**
  - Requires client-side resolution

Shared endpoint access rules (ACLs)

- **Administrator role required to delegate access managers**
- **Access manager role required to manage permission/ACLs**
- **Operations:**
  - Get list of access rules
  - Get access rule by id
  - Create access rule
  - Update access rule
  - Delete access rule

Management API

- **Allow endpoint administrators to monitor and manage all tasks with endpoint**
  - Task API is essentially the same as for users
  - Information limited to what they could see locally
- **Cancel tasks**
- **Manage pause rules**

docs.globus.org

Walk-through Jupyter Notebook

- Jupyter (iPython) notebook demonstrating use of Python SDK

github.com/globus/globus-jupyter-notebooks

- Overview
- Open source, enjoy

Walk-through

Jupyter Notebook

Exercise: Jupyter notebook

Install Jupyter notebook either locally or on EC2 instance

github.com/globus/globus-jupyter-notebooks.git

Modify Jupyter notebook to:

1. Find the endpoint id for XSEDE Comet
2. Set all the metadata fields on your shared endpoint
3. Set permissions to allow a colleague to access your shared endpoint
4. Transfer all files *.txt from the tourexercise directory on the Globus Vault endpoint to any other endpoint.
5. Monitor for completion, and monitor the event log
6. Perform an ‘ls’ given a bookmark name
7. Perform a transfer akin to ‘rsync –av –delete’
8. Anything else you want to try out...

Our Business Model
How we do what we do and how we propose to continue doing it.

RMACC 2017

August 17, 2017

Greg Nawrocki
greg@globus.org

Thank you to our sponsors

U.S. DEPARTMENT OF
ENERGY

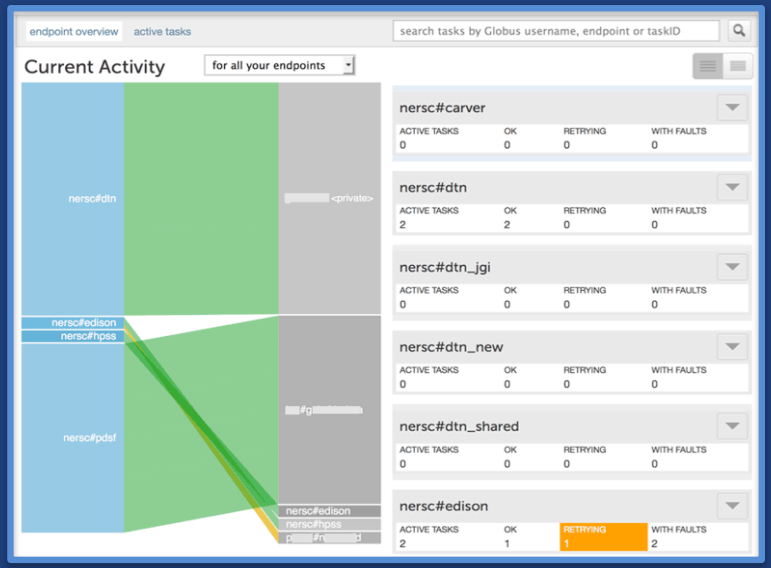
NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

THE UNIVERSITY OF
CHICAGO

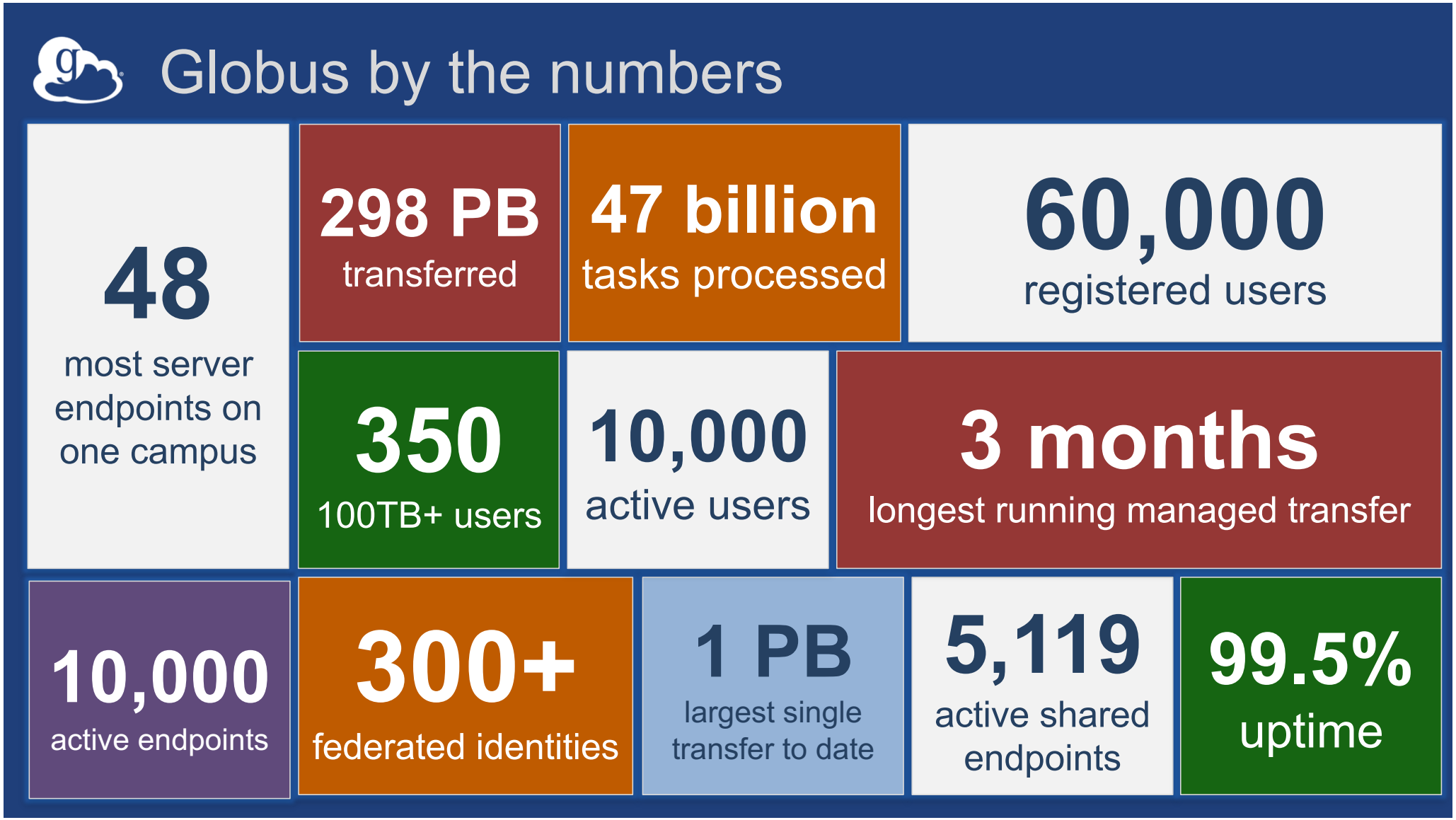
Argonne
NATIONAL LABORATORY

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
298 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!

NERSC
NIH
Berkeley UNIVERSITY OF CALIFORNIA
JOHNS HOPKINS UNIVERSITY
NCAR
UNIVERSITY OF MICHIGAN
wellcome trust sanger institute
UF UNIVERSITY of FLORIDA
Yale
CORNELL UNIVERSITY
NCSA
NEW YORK UNIVERSITY
THE UNIVERSITY OF CHICAGO
VirginiaTech Invent the Future
OAK RIDGE National Laboratory
INDIANA UNIVERSITY
MICHIGAN STATE UNIVERSITY
NEW YORK GENOME CENTER
syngenta
Dartmouth
SIMONS FOUNDATION
NIST
Los Alamos NATIONAL LABORATORY EST.1943
Argonne NATIONAL LABORATORY

Reference Documentation and Links

RMACC 2017

August 17, 2017

Greg Nawrocki
greg@globus.org

Join the Globus Community

- **Documentation**
  - docs.globus.org
- **Join the mailing lists**
  - globus.org/mailing-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.

Broadening Access to Cyberinfrastructure with Globus for Research Data Management and the Globus Platform

RMACC 2017

August 17, 2017

Greg Nawrocki
greg@globus.org