Globus for Research Data Management

Rachana Ananthakrishnan
Product Manager, Globus
ranantha@uchicago.edu

August 13 2015
Globus delivers...

(big) data transfer, sharing, publication, and discovery...

...directly from your own storage systems
“I need to easily, quickly, & reliably move or replicate my data to other systems.”
“I need to get data from a scientific instrument to my analysis system.”
“I need to easily and securely share my data with my colleagues at other institutions.”
“I need a good place to store / backup / archive my (big) research data, at a reasonable price.”
“I need to publish my data so that others can find it and use it.”
Reliable, secure, high-performance file transfer and replication

- “Fire-and-forget” transfers
- Automatic fault recovery
- Seamless security integration
- Powerful GUI and APIs

User initiates transfer request → Globus moves and syncs files → User notifies user
Amazon S3 Endpoints
Demonstration:
Sign Up
File Transfer
Simple, secure *sharing* off existing storage systems

- Easily share large data with any user or group
- No cloud storage required

1. User A selects file(s) to share, selects user or group, and sets permissions
2. Globus tracks shared files; no need to move files to cloud storage!
3. User B logs in to Globus and accesses shared file
Demonstration: Globus sharing Group Management
Curated publication of data, with relevant metadata for discovery

- Identify
- Describe
- Curate
- Verify
- Access
- Preserve

1. Researcher assembles data set; describes it using metadata (Dublin core and domain-specific)
2. Curator reviews and approves; data set published on campus or other storage
3. Peers, public search and discover data sets; transfer using Globus
Typical publication workflow

1. Publish Data
2. Describe Submission
3. Assemble Dataset (Transfer Data)
4. Curate Dataset
5. Search
6. Download

Scientist

Argonne Curator

Argonne Storage System

Globus Publish
Globus Publish enables easy and efficient publishing of large datasets. It uses Globus as a mechanism to transfer and share data.
Globus Platform-as-a-Service

- Data Publication & Discovery
- File Sharing
- File Transfer & Replication
- Identity, Group, and Profile Management
- Globus Toolkit
Platform adopters
<table>
<thead>
<tr>
<th>Globus at a glance</th>
<th>4</th>
<th>100 PB</th>
<th>20 billion</th>
<th>25,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>major services</td>
<td>13</td>
<td>transferred</td>
<td>files processed</td>
<td>registered users</td>
</tr>
<tr>
<td>national labs use Globus services</td>
<td>8,000</td>
<td>active endpoints</td>
<td>&gt;300 users are active daily</td>
<td>99.95% uptime over the past two years</td>
</tr>
<tr>
<td>&gt;30 institutions subscribed</td>
<td>1 PB</td>
<td>largest single transfer to date</td>
<td>3 months longest continuously managed transfer</td>
<td>157 federated campus identities</td>
</tr>
</tbody>
</table>
Globus endpoints worldwide
Globus is enabling…

Study of the structure and evolution of galaxies, the nature of dark energy, and cosmological history of the universe

Joel Brownstein
University of Utah

Sloan Digital Sky Survey
Source: University of Utah
Globus is enabling...

Development of numerical simulations of severe storms for improved responsiveness to weather events

Ann Syrowski
University of Illinois
Dan Kozak (Caltech) replicates 1 PB LIGO astronomy data for resilience
Erin Miller (PNNL) collects data at Advanced Photon Source, renders at PNNL, and views at ANL

Credit: Kerstin Kleese-van Dam
Globus Subscriptions

• **Provider**
  – Host shared endpoints
  – Management console
  – Data publication*
  – Amazon S3 endpoints
  – Usage reports
  – Priority support
  – Application integration

• **Branded Web Site**

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

• **Mass Storage System optimization**

  [globus.org/provider-plans](http://globus.org/provider-plans)
Questions

More information:
www.globus.org
www.globus.org/data-publication

Contact:
support@globus.org
Hands-on exercise

http://tinyurl.com/globusatpesc