Galaxy and Globus
Science as a Service

Ravi K Madduri, University of Chicago and Argonne National Laboratory
madduri@uchicago.edu
@madduri
Our vision for a 21st century discovery infrastructure

Provide more capability for more people at lower cost by delivering “Science as a service”

www.globus.org
An Old Idea:
Service Oriented Science
People **create** services (data or functions) …
which I **discover** (& decide whether to use) …
& **compose** to create a new function …
& then **publish** as a new service.

→ I find “someone else” to **host** services,
so I don’t have to become an expert in
operating services & computers!

→ I hope that this “someone else” can **manage**
security, reliability, scalability, …

“Service-Oriented Science”, *Science*, 2005
Two Broader Themes

• **Productivity** of Researchers
  – Time spent performing administrative tasks
    Vs time spent doing science
  – Reproducibility

• **Sustainability** of scientific software
  – Reduction in funding for science
NIH Grant Application Success Rates
FY 1978-2013

Time-consuming tasks in science

- Run experiments
- Collect data
- Manage data
- Move data
- Acquire computers
- Analyze data
- Run simulations
- Compare experiment with simulation
- Search the literature

- Communicate with colleagues
- Publish papers
- Find, configure, install relevant software
- Find, access, analyze relevant data
- Order supplies
- Write proposals
- Write reports
Our Science Stack

- **Galaxy**
  - Interactive execution
  - Creation, Execution, Sharing, Discovering Workflows

- **Globus**
  - Data management
  - Identity Management

- **AWS**
  - EC2, EBS, S3, SNS,
  - Spot, Route 53, Cloud Formation
Globus: Fast, reliable data transfer

1. User initiates transfer request
2. Globus moves and syncs files
3. Globus notifies user
Amazon S3 Endpoints
Globus: Sharing off existing systems

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 into Globus and accesses shared file
Globus Transfer

>25,000 registered users; >150 daily
50 PB moved; >1B files
10x (or better) performance vs. scp
99.9% availability
Entirely hosted on Amazon
Globus: Data publication service

Collection

Dataset
- Data
  - Metadata

Dataset
- Data
  - Metadata

Policies

- Metadata
- Access Control
- License
- Storage
- Curation Workflow

Community
### User Information

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>User</td>
<td>jxtx</td>
</tr>
<tr>
<td>Email</td>
<td><a href="mailto:james@taylorlab.org">james@taylorlab.org</a></td>
</tr>
<tr>
<td>Created</td>
<td>2013-12-14 19:29:48</td>
</tr>
<tr>
<td>Job Count</td>
<td>30</td>
</tr>
<tr>
<td>Bytes Transferred</td>
<td>45644557967406 (45.645 TB)</td>
</tr>
<tr>
<td>Files</td>
<td>10538680</td>
</tr>
<tr>
<td>Files Skipped</td>
<td>4175915</td>
</tr>
<tr>
<td>DN</td>
<td>/DC=org/DC=cilogon/C=US/O=Johns Hopkins/CN=James Taylor A12216</td>
</tr>
<tr>
<td>DN</td>
<td>/C=US/O=Pittsburgh Supercomputing Center/CN=James Taylor/CN=1717207743</td>
</tr>
<tr>
<td>SSH Public Key</td>
<td>ssh-dss AAAAB3NzaC1kc3MAAACBAOSZONptSidEU3Uy4B3UvLWwyda4q65bXtK0zOtzwWovh5hQSG/RWCfUIUvUGvQeVrlK34iU...</td>
</tr>
<tr>
<td>Endpoints</td>
<td>Link</td>
</tr>
</tbody>
</table>

### Active Jobs

<table>
<thead>
<tr>
<th>Status</th>
<th>User</th>
<th>Req. (UTC)</th>
<th>Tasks</th>
<th>OK</th>
<th>Failed</th>
<th>BytesTX</th>
<th>Mbps</th>
<th>Retrying</th>
<th>Duration</th>
<th>Deadline</th>
<th>Faults</th>
<th>Source</th>
<th>Destination</th>
<th>Type</th>
<th>Flags</th>
<th>Task ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONNECT_FAILED</td>
<td>jxtx</td>
<td>06-26 15:24</td>
<td>106488</td>
<td>106439</td>
<td>0</td>
<td>1.745 TB</td>
<td>30.01</td>
<td>49</td>
<td>129:13:57</td>
<td>67:46:30</td>
<td>964</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>VERIFY</td>
<td>02ead438-fd46-11e3-b575-123139403</td>
</tr>
</tbody>
</table>

### Job History

<table>
<thead>
<tr>
<th>User</th>
<th>Completed (UTC)</th>
<th>Tasks</th>
<th>OK</th>
<th>Failed</th>
<th>BytesTX</th>
<th>Mbps</th>
<th>Duration</th>
<th>Faults</th>
<th>Source</th>
<th>Destination</th>
<th>Type</th>
<th>Flags</th>
<th>Task ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>jxtx</td>
<td>2014-07-01 20:20:36</td>
<td>83469</td>
<td>83469</td>
<td>0</td>
<td>673.296 GB</td>
<td>11.98</td>
<td>124:55:59</td>
<td>822</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>VERIFY</td>
<td>fbe6b6bc-fd45-11e3-b575-123139403</td>
</tr>
<tr>
<td>jxtx</td>
<td>2014-06-26 15:05:07</td>
<td>380204</td>
<td>380204</td>
<td>0</td>
<td>42.262 GB</td>
<td>2.29</td>
<td>41:02:44</td>
<td>370</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>VERIFY, MTIME</td>
<td>380fa2bc-fheb-11e3-b574-123139403</td>
</tr>
<tr>
<td>jxtx</td>
<td>2014-06-24 21:42:14</td>
<td>260190</td>
<td>260190</td>
<td>0</td>
<td>7.709 TB</td>
<td>384.15</td>
<td>44:35:41</td>
<td>7</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>SYNC=3, VERIFY</td>
<td>9db1e742-fa72-11e3-b570-123139403</td>
</tr>
<tr>
<td>jxtx</td>
<td>2014-06-23 01:03:04</td>
<td>8664</td>
<td>8664</td>
<td>0</td>
<td>547.805 GB</td>
<td>435.54</td>
<td>02:47:42</td>
<td>0</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>SYNC=3, VERIFY</td>
<td>b3fa0c06-fa5a-11e3-b56d-123139403</td>
</tr>
<tr>
<td>jxtx</td>
<td>2014-06-21 18:21:23</td>
<td>272732</td>
<td>272732</td>
<td>0</td>
<td>1.773 TB</td>
<td>150.46</td>
<td>26:11:06</td>
<td>3</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>SYNC=3, VERIFY</td>
<td>5e1c6ff9-f895-11e3-b56b-123139403</td>
</tr>
<tr>
<td>jxtx</td>
<td>2014-06-20 16:08:34</td>
<td>856561</td>
<td>856561</td>
<td>0</td>
<td>246.727 KB</td>
<td>0.00</td>
<td>04:07:12</td>
<td>0</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>SYNC=3, VERIFY, MTIME</td>
<td>98e6a9a6-f872-11e3-b56b-123139403</td>
</tr>
<tr>
<td>jxtx</td>
<td>2014-06-20 04:12:10</td>
<td>78312</td>
<td>78312</td>
<td>0</td>
<td>373.587 GB</td>
<td>254.77</td>
<td>03:15:31</td>
<td>2</td>
<td>jxtx#rhyolite</td>
<td>vincent#dslogin01</td>
<td>Web</td>
<td>SYNC=3, VERIFY, MTIME</td>
<td>bc5c1b8b-f015-11e3-b56b-123139403</td>
</tr>
</tbody>
</table>
Flexible, scalable, affordable genomics analysis for all biologists
Globus Genomics

Globus Provides a high-performance file transfer service between all data-endpoints.

- Fault-tolerant
- Secure file transfer service between all data-endpoints

Data Management

Data Analysis

Galaxy Based Workflow Management System

- Galaxy Integrated with Globus Genomics
- Web-based UI
- Drag-Drop workflow creations
- Easily modify workflows with new tools
- Analytical tools are automatically run on the scalable compute resources when possible

Globus Genomics on Amazon EC2

globus.org/genomics
Core Capabilities

• Computational profiles for various analysis tools to provide optimal performance
• Resources can be provisioned on-demand with Amazon Web Services cloud based infrastructure
• High performance, Reliable Data movement is streamlined with integrated Globus file-transfer functionality
• Integrated Globus endpoints and Campus login
Diversity of collaborations
Usage has been promising

2.5 Million Core hours
FACE-IT (Globus + Galaxy -> Climate)
Cardio Vascular Research Grid – Dr. Winslow @ Hopkins
## Affordability/TCO

<table>
<thead>
<tr>
<th></th>
<th>Exome</th>
<th>Whole Genome</th>
<th>RNA-Seq.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$5 - $30</td>
<td>$20 - $100</td>
<td>$5 - $10</td>
</tr>
</tbody>
</table>

- Pricing based on example of paired-end fastq files with 5 Gbases.
- Pipeline includes quality control, alignment, variant calling, and annotation using the GATK best-practices pipeline.

- Pricing based on example of paired-end fastq files with 80 Gbases.
- Pipeline includes quality control, alignment, variant calling, and annotation.

- Pricing based on example of paired-end fastq files with 5 Gbases.
- Pipeline includes quality control, alignment, exon count using cufflinks, and HT-Seq count.

- **Pricing includes**
  - Estimated compute
  - Storage (one month)
  - Globus Genomics platform usage
  - Support
Sustainability

• Our goal is to build service that lives beyond a funded proposal
• Two pricing options and multiple usage tiers.
  – Targeted users include individual research groups and bioinformatics cores
  – Platform pricing (includes only subscription to the Globus Genomics platform)
  – Bundled pricing (includes Globus Genomics platform subscription and AWS usage costs)
• More information on Globus Genomics and to sign up for a free trial:
  www.globus.org/genomics
• More information on Globus: www.globus.org
• We are hiring!
Our work is supported by:
Thank you!

@madduri