Big process for big data

Process automation for data-driven science

Ian Foster
Computation Institute
Argonne National Laboratory & The University of Chicago

Talk at Astroinformatics 2012, Redmond, September 10, 2012
Researchers struggle with data

More data, more complex data
Ad-hoc solutions
Inadequate software, hardware
Data plan mandates
Complexity is large and growing

Time

Run experiment
Collect data
Move data
Check data
Annotate data
Share data
Find similar data
Link to literature
Analyze data
Publish data
"Well, in our country," said Alice ...
"you'd generally get to somewhere else — if you run very fast for a long time, as we've been doing."

"A slow sort of country!" said the Queen. "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that!"
Arlington, VA, June 2012
Jun 10 - Jun 14, 2012 - Arlington, VA
Arlington, VA; Boston, MA

Travelers
Ian Foster

Non-Travelers
Brigitte Raumann, asmyth1

Visible to
Tripl connections, Tripl Groups

Who's close
Daniel S Katz, Jennifer M Schopf (+3 others)

Trip Description
Add a Description

Trip Cost: $1,736.10

Offers for Your Trip
- Cleveland Park: $35 for a deep tissue massage at Facials by Camille
- Wine Tasting Pedicab Tour For Two People
- One ($24) or 12 ($300) Tickets to North End Pizza Tour
- One ($39) or Two ($69) Acupuncture Sessions with Consultation
- Entry for One, Two, or Four to the CitySolve Urban Race on…

Itinerary: Expand | Collapse

Sun, Jun 10
Boston, MA - Avg: Hi 79°F / Lo 57°F

4:03 PM CDT

Chicago (ORD) to Boston (BOS)
United Airlines 349 - Conf # NZVRFM
Aircraft Airbus A319
nonstop 2h, 21m 864 mi E
Purchase

Depart: Chicago (ORD), 4:03pm CDT, terminal 1, gate C17
Arrive: Boston (BOS), 7:22pm EDT (orig arr time: 7:24pm), terminal C, gate C17

Passenger
Ian Foster FF #ABL3XXXX Ticket
#0162328680876

Booking Information
Booked on United 5/25/2012
http://www.united.com/
Tripit exemplifies process automation.
Software as a Service (Gartner)

1. The application is owned, delivered, and managed remotely by one or more providers
2. The application is based on a single code base that is consumed in a one-to-many model by all contracted customers at any time
3. The application is licensed on pay-per-use or subscription basis
4. The application behind the service is properly web architected—not an existing application web enabled [D. Terrar]
Complexity is large and growing

Run experiment
Collect data
Move data
Check data
Annotate data
Share data
Find similar data
Link to literature
Analyze data
Publish data
Process automation for science

Run experiment
Collect data
Move data
Check data
Annotate data
Share data
Find similar data
Link to literature
Analyze data
Publish data

Time

Research IT as a service
Process automation for science

Run experiment
Collect data
Move data
Check data
Annotate data
Share data
Find similar data
Link to literature
Analyze data
Publish data

Time

Research IT as a service
Eppur si muove ("and yet it moves") ...

Galileo said this about the earth (the earth is not the center of the universe)

Same observation holds about data, which is often in the "wrong place" for various reasons
Reliable, high-performance, secure file transfer. Move files fast. No IT required.

WATCH A VIDEO
Globus Online in a nutshell

GET STARTED
Sign up and get moving

6723492472 MB TRANSFERRED

Why Use Globus Online?
See how easy file transfer can be

For HPC Resource Owners
Enable Globus Online for your users

For Developers
Integrate with Globus Online
Globus Transfer details

- In 20 months
  - 6,000 users
  - 6 PB moved
  - 500M files
  - 99.9% uptime

- Broad adoption
  - Experimental facilities
  - Supercomputers
  - Campuses
  - Individuals
  - Projects
<table>
<thead>
<tr>
<th>Status</th>
<th>Label</th>
<th>Task Progress</th>
<th>Completion Time</th>
<th>Request Time</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Task ld:5cbcf6bc-df3e-11e1-bf56-1231380b8963</td>
<td>197 / 1026</td>
<td>08/06/2012 03:34 AM</td>
<td>08/05/2012 11:46 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:1fe43dd4-df1d-11e1-bf56-1231380b8963</td>
<td>4362 / 4362</td>
<td>08/05/2012 11:45 AM</td>
<td>08/05/2012 11:44 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:b3c33db8-defe-11e1-bf56-1231380b8963</td>
<td>0 / 1 / 1</td>
<td>08/05/2012 11:42 AM</td>
<td>08/05/2012 08:08 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:a37697c4-defe-11e1-bf56-1231380b8963</td>
<td>624 / 624</td>
<td>08/05/2012 10:08 AM</td>
<td>08/05/2012 08:08 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:d370fd7c-deac-11e1-bf56-1231380b8963</td>
<td>3738 / 3738</td>
<td>08/05/2012 07:51 AM</td>
<td>08/04/2012 10:22 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:278af9f2-dea9-11e1-bf56-1231380b8963</td>
<td>624 / 624</td>
<td>08/05/2012 12:12 AM</td>
<td>08/04/2012 09:56 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:6d8aa43c-dcfd-11e1-bf56-1231380b8963</td>
<td>299 / 717 / 1016</td>
<td>08/02/2012 07:00 PM</td>
<td>08/02/2012 06:54 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:c150b7e-decf-11e1-bf56-1231380b8963</td>
<td>56 / 528 / 624</td>
<td>08/02/2012 06:49 PM</td>
<td>08/02/2012 06:42 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:e84f6718-dc66-11e1-bf56-1231380b8963</td>
<td>11 / 11</td>
<td>08/02/2012 06:08 PM</td>
<td>08/02/2012 06:08 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:784d8008-dc66-11e1-bf56-1231380b8963</td>
<td>11 / 11</td>
<td>08/02/2012 06:04 PM</td>
<td>08/02/2012 06:04 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:33f4206a-dc66-11e1-bf56-1231380b8963</td>
<td>11 / 11</td>
<td>08/02/2012 06:03 PM</td>
<td>08/02/2012 06:02 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:60950ae-dc66-11e1-bf56-1231380b8963</td>
<td>9 / 2 / 11</td>
<td>08/02/2012 06:01 PM</td>
<td>08/02/2012 05:57 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:a4273bfe-dc00-11e1-bf56-1231380b8963</td>
<td>11 / 11</td>
<td>08/02/2012 09:13 PM</td>
<td>08/02/2012 05:23 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:77993940-d2c4-11e1-bf56-1231380b8963</td>
<td>1339 / 1339</td>
<td>07/21/2012 06:18 PM</td>
<td>07/20/2012 06:41 PM</td>
</tr>
<tr>
<td></td>
<td>Task ld:be22047c-d0ea-11e1-bf56-1231380b8963</td>
<td>1078 / 1078</td>
<td>07/18/2012 10:17 AM</td>
<td>07/18/2012 10:16 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:d927feec-d0e6-11e1-bf56-1231380b8963</td>
<td>1077 / 1077</td>
<td>07/18/2012 10:13 AM</td>
<td>07/18/2012 10:11 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:19b4e54a-d0da-11e1-bf56-1231380b8963</td>
<td>1 / 1</td>
<td>07/18/2012 10:14 AM</td>
<td>07/18/2012 09:44 AM</td>
</tr>
<tr>
<td></td>
<td>Sync test</td>
<td>1078 / 1078</td>
<td>07/18/2012 10:17 AM</td>
<td>07/18/2012 10:16 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:be22047c-d0ea-11e1-bf56-1231380b8963</td>
<td>1077 / 1077</td>
<td>07/18/2012 10:13 AM</td>
<td>07/18/2012 10:11 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:d927feec-d0e6-11e1-bf56-1231380b8963</td>
<td>1 / 1</td>
<td>07/18/2012 10:14 AM</td>
<td>07/18/2012 09:44 AM</td>
</tr>
<tr>
<td></td>
<td>Task ld:19b4e54a-d0da-11e1-bf56-1231380b8963</td>
<td>1 / 1</td>
<td>07/18/2012 08:12 AM</td>
<td>07/18/2012 08:11 AM</td>
</tr>
</tbody>
</table>
### Transfer Activity

#### Transfer Details

- **Task ID:** 5cbcf6bc-dfe3-11e1-bf56-1231380b8963
- **Status:** ACTIVE
- **Origin:** isimipgo#us-archive
- **Destination:** xseed#trestles
- **User:** ian
- **Directories:** 10
- **Bytes Transferred:** 1853448810
- **Files:** 3731

#### Task Statistics

- **Pending:** 3342
- **Cancelled:** 0
- **Failed:** 0
- **Succeeded:** 399
- **Expired:** 0
- **Retrying:** 0

#### Completion Time

- **Request Time:** 08/06/2012 11:25 AM
- **Deadline:** 08/07/2012 11:25 AM
- **Completion Time:**
- **11/11**
- **1339/1339**
- **1556/1556**
- **1078/1078**
- **1077/1077**
Globus Online Notification
To: Ian Foster

Task a37697c4-defe-11e1-bf56-1231380b8963: SUCCEEDED

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task ID</td>
<td>a37697c4-defe-11e1-bf56-1231380b8963</td>
</tr>
<tr>
<td>Task Type</td>
<td>TRANSFER</td>
</tr>
<tr>
<td>Parent Task ID</td>
<td>n/a</td>
</tr>
<tr>
<td>Status</td>
<td>SUCCEEDED</td>
</tr>
<tr>
<td>Request Time</td>
<td>2012-08-05 13:08:24Z</td>
</tr>
<tr>
<td>Deadline</td>
<td>2012-08-08 14:10:32Z</td>
</tr>
<tr>
<td>Completion Time</td>
<td>2012-08-05 15:08:31Z</td>
</tr>
<tr>
<td>Total Tasks</td>
<td>624</td>
</tr>
<tr>
<td>Tasks Successful</td>
<td>624</td>
</tr>
<tr>
<td>Tasks Expired</td>
<td>0</td>
</tr>
<tr>
<td>Tasks Canceled</td>
<td>0</td>
</tr>
<tr>
<td>Tasks Failed</td>
<td>0</td>
</tr>
<tr>
<td>Tasks Pending</td>
<td>0</td>
</tr>
<tr>
<td>Tasks Retrying</td>
<td>0</td>
</tr>
<tr>
<td>Command</td>
<td>API 0.10 GO</td>
</tr>
<tr>
<td>Label</td>
<td>n/a</td>
</tr>
<tr>
<td>Data Encryption</td>
<td>No</td>
</tr>
<tr>
<td>Checksum Verification</td>
<td>No</td>
</tr>
<tr>
<td>Sync Level</td>
<td>0</td>
</tr>
<tr>
<td>Files</td>
<td>623</td>
</tr>
<tr>
<td>Files Skipped</td>
<td>0</td>
</tr>
<tr>
<td>Directories</td>
<td>1</td>
</tr>
<tr>
<td>Bytes Transferred</td>
<td>440658434897</td>
</tr>
<tr>
<td>Bytes checksummed</td>
<td>0</td>
</tr>
<tr>
<td>MBits/sec</td>
<td>489.213</td>
</tr>
<tr>
<td>Faults</td>
<td>1</td>
</tr>
</tbody>
</table>
Dark Energy Survey

- Receives 100,000 files each night in Illinois
- Transmit files to Texas for analysis ... then move results back to Illinois
- Process must be reliable, routine, and efficient
- Use of Globus Transfer avoids need for custom software solution

Image credit: Roger Smith/NOAO/AURA/NSF
Blue Waters has partnered with the Globus Online file transfer service.

You may access this service by entering your Blue Waters username and password.

NOTE - If you are accessing this file transfer service for the first time, you will be asked to link your Blue Waters account to a Globus Online account (if you don't have a Globus Online account you'll be able to create one).
Globus Transfer under the covers

1. User initiates transfer
   - Web GUI or command line interface

2. Globus Online moves files

3. Globus Online notifies user

Source → Globus Online → Destination
Globus Transfer under the covers

1. User initiates transfer
2. Globus Online moves files
3. Globus Online notifies user

Identities, profiles, groups
Replicated
Cloud-hosted
Web servers, data movers, CLI, etc.
High availability
Transfer state

Source
Destination

Globus Nexus
Globus Transfer
Moving beyond data movement

Dark Energy Survey      Metagenomics      Climate science
Genomics               Land use change   X-ray source data
Biomedical imaging     High energy physics Nielsen data
Moving beyond data movement

Research Data Management-as-a-Service

Globus Transfer  Globus Storage  Globus Collaborate  Globus Catalog  ...SaaS

Globus Integrate  ...PaaS

(1) Collect Data
(2) Move to Storage Store
(3) Ingest Processing
(4) Move to Community Store
(5) Publish in Registry

(6) Validate
(7) Backup
(8) Mirror
(9) Search, Browse, Analyze, Update, Annotate

User

Archival storage
Secondary site
Process automation for science

- Run experiment
- Collect data
- Move data
- Check data
- Annotate data
- Share data
- Find similar data
- Link to literature
- Analyze data
- Publish data

Research IT as a service

Time
Globus Storage: For when you want to ... 

- **Place** your data where you want
- **Access** it from anywhere via different protocols
- Update it, version it, and take snapshots
- **Share** versions with who you want
- **Synchronize** among locations
Globus Storage under the covers

Conventional or cloud storage system

Data

File system metadata

GridFTP server

HTTP server

GridFTP server

Globus Nexus
Globus Collaborate: For when you want to

Join with a few or many people to:

- Share documents
- Track tasks
- Send email
- Share data
- Do whatever

With:

- Common groups
- Delegated management
Globus Collaborate under the covers

Collaborative tool

Identity, profile, group database hosted on Amazon
Data acquisition, management, analysis

Big Data (volume, velocity, variety, variability) ... demands Big Process in order for discovery to scale
Let’s rethink how we provide research IT

Accelerate discovery and innovation worldwide by providing **research IT as a service**

Leverage the cloud to

- provide millions of researchers with unprecedented access to powerful tools;
- enable a massive shortening of cycle times in time-consuming research processes; and
- reduce research IT costs dramatically via economies of scale
MG-RAST (the Metagenomics RAST) server is an automated analysis platform for metagenomes providing quantitative insights into microbial populations based on sequence data.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of metagenomes</td>
<td>40,895</td>
</tr>
<tr>
<td># base pairs</td>
<td>11.22 Tbp</td>
</tr>
<tr>
<td># of sequences</td>
<td>104.3 billion</td>
</tr>
<tr>
<td># of public metagenomes</td>
<td>7,391</td>
</tr>
</tbody>
</table>

The server provides web based upload, quality control, automated annotation and analysis for samples up to 10GBp. Comparison between large numbers of samples is enabled via pre-computed abundance profiles.

This project has been funded in part with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. HHSN272200900040C.
**United States Wind Energy Potential Test.csv**

*Click here to add a description*

```sql
SELECT * FROM [chartest].[table_Untited States Wind Energy Potential.csv]
```

### Dataset Preview

(rows 1 - 48 of 48)

<table>
<thead>
<tr>
<th>State</th>
<th>Total (km)</th>
<th>Excluded (km)</th>
<th>Available (km)</th>
<th>Available % of State</th>
<th>% of Total Windy Land Excluded</th>
<th>Potential Installed Capacity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alabama</td>
<td>80.36</td>
<td>56.72</td>
<td>23.64</td>
<td>0.0176864506525858%</td>
<td>70.5823792931807%</td>
<td>118.2</td>
</tr>
<tr>
<td>Arizona</td>
<td>4544.96</td>
<td>2364.14</td>
<td>2180.82</td>
<td>0.736642763731711%</td>
<td>52.018739421249%</td>
<td>10904.1</td>
</tr>
<tr>
<td>Arkansas</td>
<td>4683.24</td>
<td>2823.18</td>
<td>1840.06</td>
<td>1.33593638826854%</td>
<td>60.541168010911%</td>
<td>9200.3</td>
</tr>
<tr>
<td>California</td>
<td>26901.28</td>
<td>20079.24</td>
<td>6622.04</td>
<td>1.66671846119056%</td>
<td>74.6404632047248%</td>
<td>34110.2</td>
</tr>
<tr>
<td>Colorado</td>
<td>95830.36</td>
<td>18386.46</td>
<td>77443.9</td>
<td>28.7253372095834%</td>
<td>19.1854665853284%</td>
<td>387219.5</td>
</tr>
<tr>
<td>Connecticut</td>
<td>31.36</td>
<td>26.00</td>
<td>5.3</td>
<td>0.0413651433887587%</td>
<td>83.0994897959184%</td>
<td>26.5</td>
</tr>
<tr>
<td>Delaware</td>
<td>36.56</td>
<td>34.66</td>
<td>1.9</td>
<td>0.03739688338729%</td>
<td>94.8030634573034%</td>
<td>9.5</td>
</tr>
<tr>
<td>Florida</td>
<td>9.56</td>
<td>4.98</td>
<td>4.53</td>
<td>0.0000548155835831%</td>
<td>99.163179916318%</td>
<td>0.4</td>
</tr>
<tr>
<td>Georgia</td>
<td>281.28</td>
<td>255.26</td>
<td>26.02</td>
<td>0.0170827943529%</td>
<td>90.749431177861%</td>
<td>130.1</td>
</tr>
<tr>
<td>Idaho</td>
<td>13420.4</td>
<td>9805.28</td>
<td>3615.12</td>
<td>1.67025347276445%</td>
<td>73.0625018628357%</td>
<td>18075.6</td>
</tr>
<tr>
<td>Illinois</td>
<td>70763.66</td>
<td>20787.14</td>
<td>49976.42</td>
<td>34.2484872158604%</td>
<td>29.3754867491902%</td>
<td>249882.1</td>
</tr>
<tr>
<td>Indiana</td>
<td>4825.24</td>
<td>18620.74</td>
<td>2945.5</td>
<td>0.24622902037802%</td>
<td>35.0938907416847%</td>
<td>149227.6</td>
</tr>
</tbody>
</table>
Acknowledgements

• Thanks for vital and much appreciated support:
  – NSF Office of Cyberinfrastructure (OCI)
  – DOE Office of Advanced Scientific Computing Research (ASCR)
  – National Institutes of Health
  – The University of Chicago

• Thanks to the amazing Globus Online team at the University of Chicago and Argonne. See www.globusonline.org/about/goteam/
Thank you!

globusonline.org
@globusonline

foster@anl.gov
foster@uchicago.edu