Big process for big data
Process automation for data-driven science

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

Talk at NIST Big Data Workshop, Gaithersburg, June 13, 2012
Big science has achieved big successes

LIGO: 1 PB data in last science run, distributed worldwide

Robust production solutions
Substantial teams and expense
Sustained, multi-year effort
Application-specific solutions, built on common technology

OSG: 1.4M CPU-hours/day, >90 sites, >3000 users, >260 pubs in 2010

ESG: 1.2 PB climate data delivered to 23,000 users; 600+ pubs

All build on NSF- & DOE-supported Globus Toolkit software
But small science is struggling

More data, more complex data
Ad-hoc solutions
Inadequate software, hardware
Data plan mandates
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
A first take on “big process for science”

Dark Energy Survey   Metagenomics   Climate science
Genomics   Land use change   X-ray source data
Biomedical imaging   High energy physics   Nielsen data
A first take on “big process for science”

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

(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
Registry
Community store
Ingest store
Temporary Storage
Archival storage
Secondary site
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

4,194,849,260 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 18 months
  - 5,000 users
  - 5 PB
  - 500M files
  - 99.9% uptime

- Broad adoption
  - Experimental facilities
  - Supercomputers
  - Campuses
  - Individuals
  - Projects
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).
A first take on “big process for science”

Dark Energy Survey  Metagenomics  Climate science
Genomics  Land use change  X-ray source data
Biomedical imaging  High energy physics  Nielsen data
A first take on “big process for science”
A need for standards and best practices

• Security of “research IT as a service” products
  – Best practices to allow their use for sensitive data

• Distributed implementation of such services
  – Franchise model to enable broader adoption
Data acquisition, management, analysis

Big Data (volume, velocity, variety, variability)
... demands Big Process in order for discovery to scale

Experiments

Literature don’t forget!

Computations
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
Acknowledgements

• Thanks for vital and much appreciated support:
  – DOE Office of Advanced Scientific Computing Research (ASCR)
  – NSF Office of Cyberinfrastructure (OCI)
  – 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