Globus for Big Data & Science Gateways @ LBNL

Krishna Muriki kmuriki@lbl.gov
Karen Fernsler kmfernsler@lbl.gov
High Performance Computing Services (HPCS), IT Division
Lawrence Berkeley National Laboratory (LBNL)

GlobusWorld
April 16th 2014
Data Volumes at LBNL & UCB

- LBNL IT Division HPCS group
  - support science
  - PI clusters
  - Institutional cluster
  - Condo Computing
  - Data transfer services
  - Web portal services

- Daily data volume estimates more than 70 TB /day.
- Advanced Light Source
  - X-ray Micro Tomography instrument - 40TB in short period.
- Physics Division
- Molecular Foundry
LBNL IT + NERSC participate in the Partner Program

- Branded websites
- Support for LBNL Identity federation
- Bind your LBNL & GO accounts once
- Use LBNL account to access GO portal (via InCommon alternate login)

- 2500 Globus Plus subscriptions
- 25 Managed endpoints
- Improved support.
• Two projects with Globus @ LBNL
  • Use of Gridftp & Globus Online for SDSS III project
  • Implementing NEWT web portal over GSI-SSH
• Large Imaging & Spectroscopic survey of the Northern Sky
  - Four data surveys: BOSS, SEGUE-2, APOGEE, & MARVELS
  - LBNL is Tier-1 landing site for telescope data
  - Daily data extraction pipeline operations at LBNL

• Baryon Oscillation Spectroscopic Survey (BOSS)
  - Survey to map the universe (PI: David Schlegel @ LBNL)
  - Goal to map 1.5 million galaxies, 150K quasars, & many stars.
• Total data volumes in the range of 200TB

• Public data access
  - Periodic data releases for research community
  - DR10 - 70TB, DR9 - 60TB, DR8 - 49.5TB

• Power user data access
  - Mirroring of total dataset to other centers
  - Daily TBs across the country (coast to coast)

• Normal cluster user data access
  - LBL cluster users transferring results back to local campuses
  - Berkeley, Univ of Utah, New York Univ, John Hopkins, etc..
• Normal Cluster User data access
  - Previously using scp or rsync on the cluster nodes

  ➢ Provided a gridftp server and a standard GO endpoint (lbnl#riemann)
  ➢ Accepts cluster OTP authentication via MyProxy Oauth.
  ➢ Proxy life time increased to 4 days

• Public data access
  - Previously made accessible as http downloads
  - Webserver serving selected data in read-only mode.

  ➢ Set up an anonymous gridftp server serving the selected data
  ➢ A GO endpoint (lbnl#sdss3) serving the data.
  ➢ Anonymous endpoint so no authentication or proxy life time.
• Power User data access
  - 4 day proxy life time of standard GO endpoint not convenient
  - GO Shared endpoints stay active forever
  - Cluster gridftp server enabled for sharing.
  - ‘lbnl’ GO account updated with Plus subscription.
  - lbnl#riemann activated with power user credential.
  - New shared GO endpoint (lbnl#riemann-share) created
  - Shared endpoint accessible by power user(s)
  - We need to better understand the security implications.

• Collaborators data access
  - http downloads & rsync supports group passwords
  - Need to implement shared endpoints with write permissions
Cluster Webportals using NEWT

- Nice and Easy Web toolkit from NERSC
- Web service to access resources via RESTful API

- Provides many resource URLs
  - /login, /logout
  - /job, /file, /queue

- Heavily relies on Globus toolkit
  - Job url via Globus Gatekeeper (GRAM) service.
  - File url via GridFTP service.

- `curl -k -c cookie.txt -X POST -d "username=XYZ&password=PASS" https://ws.hpcs.lbl.gov/newt/auth`
- `curl -k -b newt_cookies.txt -X GET https://ws.hpcs.lbl.gov/newt/file/etc/motd?view=read`
NEWT Implementation with GSI-SSH

NEWT Restful API

/Job/$ID
/Job/Path
/Queue
/Auth
/Command/

User specific code using our API
(curl or wget)

Webserver outside Supercluster firewall

HTTP cookies holding session state & short lived user proxies

Port 22, (GSSSH, GSISCP)

Port 2811 GridFTP

HPCS Cluster Infrastructure

Supercluster firewall

HPCS MyProxy Service (with OAuth)

LBL IT
Cryptocard/Nordic Edge Service
Next steps:

- Collaborators data access via group passwords
- Shared endpoints security implications

Acknowledgements

Many thanks to

- Shreyas Cholia @ NERSC, LBNL