Adapting federated cyberinfrastructure for shared data collection facilities in structural biology

GlobusWorld 2012

Ian Stokes-Rees ijstokes@seas.harvard.edu @ijstokes
Harvard Medical School
Adapting federated cyberinfrastructure for shared data collection facilities in structural biology

Ian Stokes-Rees, Ian Levesque, Frank V. Murphy IV, Wei Yang, Ashley Deacon and Piotr Sliz

JCSG (U54 GM094586 and GM074898). Portions of this research were performed at the Stanford Synchrotron Radiation Lightsource (SSRL), SLAC National Accelerator Laboratory. The SSRL is a national user facility operated by Stanford University on behalf of the US Department of Energy, Office of Basic Energy Sciences. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. The authors would also like to thank Steve Tuecke, Rachana Ananthakrishnan and Raj Kettimuthu from Globus Online for their support and encouragement with this project, and James Withrow from NE-CAT for technical support in the deployment process. Terrence Martin, at the University of California San Diego (UCSD), and Brian Bockleman, at the University of Nebraska-Lincoln, provided invaluable assistance with service configuration and data management on the HadoopFS system at UCSD.
Structural Biology: Study of Protein Structure and Function

- Shared scientific data collection facility
- Data intensive (10-100 GB/day)
Cryo Electron Microscopy

- Previously, 1-10,000 images, managed by hand
- Now, robotic systems collect millions of hi-res images
- Estimate 250,000 CPU-hours to reconstruct model
- 20-50 TB of data in most extreme case
Molecular Dynamics Simulations

1 fs time step
1 ns snapshot
1 us simulation
1e6 steps
1000 frames
10 MB / frame
10 GB / sim
20 CPU-years
3 months (wall-clock)
Boston Life Sciences Hub

- Biomedical researchers
- Government agencies
- Life sciences
- Universities
- Hospitals
Data Access
Globus Online: High Performance Reliable 3rd Party File Transfer

GUMS: MyProxy Credential Management Service
- DN to user mapping

VOMS: VO membership

Cluster portal

Certificate Authority
- root of trust

Globus Online file transfer service

Lab file server

Data collection facility

Desktop

Laptop
User can directly access lab or facility data from laptop

Local accounts within lab infrastructure

Shared (lab level) accounts at facility

Tiered storage

Public access available to archived data through web interface

Embargo policy to hold deposited data for agreed time

VO management

Tier 1
- 6 month staging storage
- /stage/sliz
- /stage/murphy

Tier 2
- 10 TB per group permanent archive
- /data/sliz
- /data/murphy
- /data/deacon

Tier 3
- 10 PB public archive
- /public/2009/
- /public/2009/
- /public/2009/
- /embarg/2010/
- /embarg/2011/

NE-CAT beamline at APS

NEBCAT beamline at APS

Harvard

"Scott" from the Sliz lab

general public

laptop

laptop

WWW

Tier 1

Tier 2

Tier 3

Globus Online

SBGrid Science Portal

MyProxy

Credential Management Service

VOMS
Architecture

✦ SBGrid
  • manages all user account creation and credential mgmt
  • hosts MyProxy, VOMS, GridFTP, and user interfaces

✦ Facility
  • knows about lab groups
    • e.g. “Harrison”, “Sliz”
  • delegates knowledge of group membership to SBGrid VOMS
    • facility can poll VOMS for list of current members
  • uses X.509 for user identification
  • deploys GridFTP server

✦ Lab group
  • designates group manager that adds/removes individuals
  • deploys GridFTP server or Globus Connect client

✦ Individual
  • username/password to access facility and lab storage
  • Globus Connect for personal GridFTP server to laptop
  • Globus Online web interface to “drive” transfers
Objective

✦ Easy to use high performance data mgmt environment
✦ Fast and reliable file transfer
  • facility-to-lab, facility-to-individual, lab-to-individual
✦ Reduced administrative overhead
✦ Better data curation
✦ Release data to public after embargo period
Challenges

- **Access control**
  - visibility
  - policies

- **Provenance**
  - data origin
  - history

- **Meta-data**
  - attributes
  - searching
Grid Computing Today
Capabilities

✦ Server-to-server interaction
✦ Federated aggregation of CPU power
✦ Predictable data patterns
✦ Command-line access from specialized servers
✦ Web-based access through “Science Gateways”
  • pre-defined computing workflows
  • e.g. SBGrid Science Portal
Weaknesses

- Federated user identity system
  - X.509 digital certificates and associated apparatus
- Ease of use for end users
- Data management tools
- Support for collaborations
Opportunities

✧ “Last mile” challenge
  • to the desktop
  • to the laptop

✧ Unified identity management
  • centralized set of credentials for each person

✧ Empower collaborations to self-manage

✧ Shift of focus from “compute” to “data”
  • for users
  • for facilities where data is the main challenge
User Credentials
X.509 Digital Certificates

✦ Analogy to a passport:
  • Application form
  • Sponsor’s attestation
  • Consular services
    • verification of application, sponsor, and accompanying identification and eligibility documents
  • Passport issuing office

✦ Portable, digital passport
  • fixed and secure user identifiers
    • name, email, home institution
  • signed by widely trusted issuer
  • time limited
  • ISO standard
Addressing Certificate Problems

- Request signed cert
- Return tracking number
- Notify agents
- Review request
- Approve cert
- Confirm eligibility
- Notify availability
- Retrieve cert
- Sign cert
- U1: Generate cert key pair
- U2a: Export signed cert key pair

Time line:

1. CA
2. RA Agent
3. Sponsor
4. User
VO (Group) Membership Registration

- Present cert to request membership by DN
- Request VO groups and roles
- Notify admin
- Verify user eligibility
- Confirm eligibility
- Approve membership, groups, and roles
- Notify
- Request VOMS AC
- Return VOMS AC
- Add AC to proxy cert
Process and Design Improvements

✦ Single web-form application
  • includes e-mail verification

✦ Centralized and connected credential management
  • FreeIPA LDAP - user directory and credential store
  • VOMS - lab, institution, and collaboration affiliations
  • MyProxy - X.509 credential store

✦ Overlap administrative roles
  • system admin
  • registration agent for certificate authority (approve X.509 request)
  • VO administrator to register group affiliations

✦ Automation
“Last Mile” and Ease of Use
Grid computing at your desk

✦ Platforms
  • Windows
  • OS X

✦ Web interfaces
  • SBGrid Science Portal

✦ One-stop-shop for account management
  • Centralized services
  • Automated processes
  • Administrator intervention
  • Support
Ryan, a postdoc in the Frank Lab at Columbia

Access NRAMM facilities securely and transfer data back to home institute

Check SBGrid for Ryan’s group membership

Verify lab membership in Frank Lab, so grant access to files

Initiate transfer at NRAMM using credential held by SBGrid

Notify user of completion

Use Globus Online to manage transfer from NRAMM back to lab

Request access to NRAMM facility

Verify SBGrid for Ryan’s group membership

Automated X.509 application

Automated Globus Online application
Architecture Diagrams
Acknowledgements

✦ Piotr Sliz

✦ SBGrid Science Portal
  • Daniel O’Donovan, Meghan Porter-Mahoney

✦ SBGrid System Administrators
  • Ian Levesque, Peter Doherty, Steve Jahl

✦ Facility Collaborators
  • Frank Murphy (NE-CAT/APS)
  • Ashley Deacon (JCSG/SLAC)

✦ Globus Online Team
  • Steve Tueke, Ian Foster, Rachana Ananthakrishnan, Raj Kettimuthu

✦ Ruth Pordes
  • Director of OSG, for championing SBGrid