Leveraging Digital Infrastructure and Innovative Software Services to Accelerate Scientific Discovery

Hervé Guy and Steve Tuecke | April 9, 2014 | 2014 Internet2 Global Summit | Denver, CO
CANARIE Research Software

Hervé Guy
herve.guy@canarie.ca
Canada’s National Research and Education Network (NREN)

A successful nationwide collaboration

Le Réseau national de recherche et d’éducation (RNRE) du Canada

Une collaboration modiale réussie
Building Tomorrow’s Digital Infrastructure Today

- Cloud Services
- Software for Research
- Canada’s National Research and Education Network (NREN)
- Federated Identity
Digital infrastructure is *more* than advanced networks.
Research Software Program

> Network-Enabled Platforms (NEP): Complete Software Platforms
> Research Platform Interface (RPI): Discreet Software Elements
Software Platforms for Research

A software platform optimized to leverage digital infrastructure for big data analysis in support of research, discovery and innovation.

A toolkit of reusable software services that accelerate the development of domain-specific research platforms.

A more efficient way to turn research data into knowledge.

Specifically designed by researchers for researchers.
CANARIE Research Software Paradigm: More Innovation, Less Time & Cost
The Power of Research Software Platforms

• Existing CBRAIN NEP (soon to be an RPI contributor) best illustrates a network-enabled platform

https://brainbrowser.cbrain.mcgill.ca/surface-viewer#
CBRAIN Demonstration

The Power of Research Software Elements

- Excellent, reusable functionality (services) are being built within research platforms
- But services are locked inside (most) platforms
  - Platforms were not designed to use or provide discreet services
Example: State-of-the-art remote resource management functionality/service
Software Elements (RPI) in Action

- CBRAIN exposes its software elements toolkit (e.g. Data Services, External Resource Monitoring, and Generic 3D Web-Viewer Service as an RPI) to the outside world
  - CBRAIN starts using service in this new format
  - Evolves platform design towards software element-based
- A new “Genetics and Genomics Analysis Platform” will use service elements instead of developing themselves
  - Less expensive platforms
  - Faster to research
  - Best-in-class services
  - Sustainable services through larger community
- Genetics Platform will in turn provide 3 software elements to the community (i.e. Virtualized “Galaxy” life science research platform that provides complex analysis pipelines to researchers without programming knowledge).
Software Elements (RPI) Demonstration
Generic 3D Web-Viewer Service (BrainBrowser)

Use Case: Ocean Networks Canada Observatory Platform

ONC makes their data an OGC-compliant and releases the data as a Service (an RPI). Example: water pressure

Chilean Earthquake and Tsunami of 27 Feb, 2010
The Disaster Recovery Platform adds OCG data compliance, and connects to the Oceans Network Canada Observatory RPI for real-time data.
The Result is a Highly Advanced Tsunami Detector
Efficient Reuse of Software Elements from Platforms

Previous Mandate: 21 Platforms Funded

New RPIs created

...leveraged by new Platforms

...spinoff additional RPIs

21

15

12

47
New Basis for Future Collaboration

- 15 elements available to the 12 new platforms
- **A total of 62** software elements will soon be available for platforms
  - Many more to choose from than initial palette
  - Added attraction for developers
  - Increased leverage possible for both research time and money
1st 15 Software Elements as Reused by 12 New Software Platforms

Imagine the collaboration possible with 62 Software Elements!
Effect of the Reuse of Software Elements on Platform Development Time
## CANARIE Research Software Portal:
https://science.canarie.ca

### Research Software Service List

Click on service name for more information.

<table>
<thead>
<tr>
<th>Service Name</th>
<th>Category</th>
<th>Status</th>
<th>Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CANARIE Reference Service (DAIR Alberta)</td>
<td>Other</td>
<td>✓</td>
<td>99.5%</td>
</tr>
<tr>
<td>CANARIE Reference Service (DAIR Quebec)</td>
<td>Other</td>
<td>✓</td>
<td>99.4%</td>
</tr>
<tr>
<td>Inocybe Technologies - Cloud manager with inter-cloud scheduling</td>
<td>Resource / Cloud Management</td>
<td>✓</td>
<td>98.7%</td>
</tr>
<tr>
<td>McGill University - BrainBrowser Surface Viewer Web Service</td>
<td>Data Visualization</td>
<td>✓</td>
<td>100.0%</td>
</tr>
<tr>
<td>Ocean Networks Canada - Sensor Observation Service</td>
<td>Sensor Management / Data Acquisition</td>
<td>✓</td>
<td>99.5%</td>
</tr>
<tr>
<td>University of Alberta - CICSTART CML Script Client</td>
<td>Resource / Cloud Management</td>
<td>✓</td>
<td>98.9%</td>
</tr>
<tr>
<td>University of Alberta - CICSTART Data Catalogue</td>
<td>Data Storage and Retrieval</td>
<td>✓</td>
<td>98.9%</td>
</tr>
<tr>
<td>University of Alberta - CICSTART File Transfer / Caching</td>
<td>Data Storage and Retrieval</td>
<td>✓</td>
<td>98.9%</td>
</tr>
<tr>
<td>University of Alberta - CICSTART User Authentication</td>
<td>User Management / Authentication</td>
<td>✓</td>
<td>98.9%</td>
</tr>
<tr>
<td>University of Alberta - CICSTART VFS</td>
<td>Data Storage and Retrieval</td>
<td>✓</td>
<td>98.9%</td>
</tr>
<tr>
<td>University of Calgary - GeoCENS RPI</td>
<td>Sensor Management / Data Acquisition</td>
<td>✓</td>
<td>98.2%</td>
</tr>
<tr>
<td>University of Calgary - Visualization Server Pool Management Service</td>
<td>Data Visualization</td>
<td>✓</td>
<td>100.0%</td>
</tr>
<tr>
<td>University of Victoria - Batch Services</td>
<td>Resource / Cloud Management</td>
<td>✓</td>
<td>98.1%</td>
</tr>
</tbody>
</table>
Service Monitoring – Reliability

**Reliability**

<table>
<thead>
<tr>
<th>Current Status</th>
<th>7 Day</th>
<th>28 Day</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️</td>
<td>99.5%</td>
<td>98.9%</td>
</tr>
</tbody>
</table>

**% Reliability for 28 days**

- **Legend:**
  - Green: 100% reliability
  - Brown: 95% to less than 100% reliability
  - Red: Less than 95% reliability
  - Gray: No data available

To view full failed validation results for this service, please click here.
Clicking on the reliability bar for a given day will open the failed validation results for that day.
A New Paradigm for Research Software Development

- Unlocks functional elements built within research Software Platforms
- Supports the culture of collaboration and reuse within scientific research community
- Maximizes efficiency by eliminating duplication of effort and investment
- Minimizes time to discovery:
  1. Browse / search software element repository
  2. Reuse existing software elements to build customized platform for research requirements
  3. Share new, discreet software elements with the global community
  4. Contribute to innovative software ecosystem

CANARIE’s toolkit of software elements is available to the research community at no cost:
https://science.canarie.ca

Interested in providing a software element?
support@science.canarie.ca
Building a Campus Data Service with Globus Software-as-a-Service

Steve Tuecke
tuecke@uchicago.edu
Campus Data Service Use Cases

• “I need a good place to store or backup my (big) research data, at a reasonable price.”

• “I need to easily, quickly, and reliably move or mirror portions of my data to other places, including my campus HPC system, lab server, desktop, laptop, XSEDE, cloud, etc.”

• “I need a way to easily and securely share my data with my colleagues at other institutions.”

• “I want to publish my data so that its available long-term.”

• “I want to archive my data in case its needed sometime in the future.”
What if the research work flow could be managed as easily as...

flickr
...our pictures

NETFLIX
...home entertainment

Gmail
...our e-mail
What makes these services great?

Great User Experience

+ Scalable (but invisible) infrastructure
We aspire to create a great user experience for research data management.

What would a “dropbox for science” look like?
Managing data should be easy …
... but it’s hard and frustrating!
What is Globus?

Big data transfer and sharing...

...with Dropbox-like simplicity...

...directly from your own storage systems
Reliable, secure, high-performance *file transfer* and *synchronization*

- “Fire-and-forget” transfers
- Automatic fault recovery
- Seamless security integration

**Diagram:**

1. User initiates transfer request
2. Globus moves and syncs files
3. Globus notifies user
Simple, secure sharing off existing storage systems

- Easily share large data with any user or group
- No cloud storage required

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 in to Globus and accesses shared file

- Easily share large data with any user or group
- No cloud storage required
Globus is SaaS

- Web, command line, and REST interfaces
- Reduced IT operational costs
- New features automatically available
- Consolidated support & troubleshooting
- Easy to add your laptop, server, cluster, supercomputer, etc. with Globus Connect
Optimizing the Campus Data Service with a Science DMZ

**Dedicated Systems for Data Transfer**
- Data Transfer Node
  - High performance
  - Configured for data transfer
  - Optimized tools: Globus

**Network Architecture**

**Science DMZ**
- Dedicated DTN location
- Proper security
- Easy to deploy - no need to redesign network
- Info: fasterdata.es.net

**Performance Testing & Measurement**

**perfSONAR**
- Enables fault isolation
- Verify correct operation
- Widely deployed in Esnet, other facilities

Info: fasterdata.es.net
Science DMZ Takes Many Forms

• Different approaches based on scope
  – Small installation for a project or two
  – Facility inside a larger institution
  – Institutional capability serving multiple departments/divisions

• Key point of concentration: High-latency path for TCP
The Data Transfer Node

- Dedicated high-performance server optimized for a data transfer
- Storage, networking, motherboard each tuned for optimal performance
- Typically mounts a campus-wide filesystem
- Globus is ideal DTN service: optimized for different workloads, configured for reliability
- Globus Connect Server can use multiple DTNs
Creating a Campus Data Service

Pick your favorite storage…

…put it in your Science DMZ with Data Transfer Nodes…

...add Globus as the user interface…

(...optionally build specialized web applications on Globus)
We are a non-profit service provider to the non-profit research community.
We are a non-profit service provider to the non-profit research community

Our challenge:
Sustainability
Globus End User Subscriptions
(for non-profit research & education)

• **Basic: Free**
  – File transfer and synchronization to/from servers
  – Create private and public endpoints
  – Access to shared endpoints created by others

• **Plus: $7/month (or $70/year)**
  – Create and manage shared endpoints
  – Peer-to-peer transfer and sharing

• **Commercial pricing also available**
Globus Provider Subscriptions

- **Managed Endpoints**: management console, usage reports, MSS optimization, shared endpoints
- Priority support
- Plus Subscriptions
- Branded Web Site
- Alternate Identity Provider
- Integration Support

Coming soon: NET+ Globus
Getting started

• Signup: globus.org/signup
• Connect: globus.org/globus-connect
• Subscribe: globus.org/provider-plans
• Learn: support.globus.org/forums
• Need help? support.globus.org
• Follow us: @globusonline