Information Services Discussion
TeraGrid ‘08

John-Paul (JP) Navarro
TeraGrid Grid Infrastructure Group (GIG)
Area Co-Director for Software Integration
University of Chicago, Argonne National Laboratory

June 2008
The TeraGrid's Information Services vision is to:

1) define a coordinated way for TeraGrid participants to publish what they offer users,
2) define a way for the TeraGrid to aggregate and index the information from all TeraGrid participants, and
3) to publish this information to the public in a form that can easily be used by other software, users, and TeraGrid service providers themselves.
High-Level Components

TeraGrid Wide Information Services

Service Provider Information Services

Clients

WS/REST HTTP GET

Apache 2.0

Cache

Tomcat WebMDS

WS MDS4

WS MDS4

WS/REST HTTP GET

Clients

WS/REST HTTP GET

WS/REST HTTP GET

WS/REST HTTP GET

WS/REST HTTP GET

WS/REST HTTP GET

WS/REST HTTP GET

WS/REST HTTP GET
High-Availability Design

TeraGrid Dynamic DNS

Server failover propagates globally in 15 minutes

Clients

info.teragrid.org

info.dyn.teragrid.org

TeraGrid Wide Information Services

Service Provider Information Services

Static paths

Dynamic paths
Information Services Design Goals

Applies Grid concepts to information publishing

1. Publishing is the responsibility of the information owner
2. Publishing is done using standard (content) schemas
3. Publishing thru standard interfaces regardless of content and where the data comes from
4. Publishing services should be available globally (subject to authentication/authorization)
5. **Information owners publish to EVERYONE, not just the TeraGrid**
6. Publishing is a grid service

Applies Grid concepts to aggregating information

1. Publishing aggregated information is done exactly like original information publishing
2. Aggregation uses standard information services interfaces to retrieve information
3. This is how a collaboration, such as the TeraGrid, aggregates participant information

Applies Grid concepts to querying information

1. Querying can use standard interfaces regardless on content
Service Provider vs TeraGrid Wide Services

Service Provider Information Services

- Content:
  - Locally owned and maintained information
  - Originates anywhere the service provider wishes

- Services:
  - 1 general purpose MDS service
  - 1 scheduling MDS service

TeraGrid Wide Information Services

- Content:
  - Aggregate/index service provider information
  - Additional central information (TGCDB, GIG operated services, ...)
  - Cached (service providers services can be down)
  - Authenticated registrations

- Services:
  - Several redundant servers (99.5% plus availability)
  - Information caching (persistence)
  - Several MDS4 services (WS/SOAP)
  - WebMDS/Tomcat, Apache 2.0, ... services (WS/REST)

Content published in: HTML, XML, CVS, ...
**WS/* (Tomcat 5.0, Apache 2.0)**

**Benefits**
- Very common web services platform
- Supports several web service interfaces (including simple)
- Supports multiple styles like REST, Web 2.0
- Can be highly scalable

**Content**
- Many formats: HTML, XHTML/XML, XML, RSS/Atom, ...

**WebMDS (Globus 4.0.x/VDT 1.7.1)**

**Benefits**
- Live MDS4 content access
- XPath support
- XSLT transforms

**Content**
- Many formats: HTML, XHTML/XML, XML, RSS/Atom

**WS/SOAP (Globus 4.0.x/VDT 1.x.y MDS4)**

**Benefits**
- Indexing, Trigger
- Registration, Publish, Subscribe
- Security/Authorization
- Robust WSRF interface

**Content**
- XML
Current Content

CTSS v4 Capability Kits
- Services
- Software

Site & Resource Cross-Reference
- Information Service identifiers
- TGCDB identifiers and descriptions

Science Gateways
- Descriptive Information

Local HPC Software Prototype
CTSS 4 Capability Kits

For each capability kit on each resource
- Current support level, and target support level
  - Development, Testing, Production
- Support organization and contact
- Inca status URL
- Multiple version of a kit with different support levels
Site & Resource Cross Reference

SiteID: ncsa.teragrid.org

organization_abbrev: NCSA
organization_name: National Center for Supercomputing Applications
amie_name: NCSA

<table>
<thead>
<tr>
<th>ResourceID</th>
<th>resource_code</th>
<th>resource_name</th>
<th>pops_name</th>
<th>grid_resource_name</th>
<th>resource_description</th>
</tr>
</thead>
<tbody>
<tr>
<td>abe.ncsa.teragrid.org</td>
<td>NCSA-ABE</td>
<td>abe.ncsa.teragrid</td>
<td>NCSA/LONI Dell PowerEdge Linux Clusters (Abe/Queenbee)</td>
<td>abe-queenbee.teragrid</td>
<td></td>
</tr>
<tr>
<td>cobalt.ncsa.teragrid.org</td>
<td>NCSA-COBALT</td>
<td>cobalt.ncsa.teragrid</td>
<td>NCSA SGI Altix (Cobalt)</td>
<td>cobalt.ncsa.teragrid</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCSA-COPPER</td>
<td>copper.ncsa.teragrid</td>
<td>NCSA IBM p690 (Copper)</td>
<td>copper.ncsa.teragrid</td>
<td></td>
</tr>
<tr>
<td>dtf.ncsa.teragrid.org</td>
<td>NCSA-DFT</td>
<td>dtf.ncsa.teragrid</td>
<td></td>
<td>dtf.ncsa.teragrid</td>
<td>NCSA DTF Cluster</td>
</tr>
<tr>
<td></td>
<td>NCSA-RADIUM</td>
<td>radium.ncsa.teragrid</td>
<td>NCSA Condor Pool (Radium)</td>
<td>radium.ncsa.teragrid</td>
<td></td>
</tr>
<tr>
<td>tungsten.ncsa.teragrid.org</td>
<td>NCSA-TUNGSTEN</td>
<td>tungsten.ncsa.teragrid</td>
<td>NCSA Xeon Linux Supercluster (Tungsten)</td>
<td>tungsten.ncsa.teragrid</td>
<td></td>
</tr>
</tbody>
</table>
Science Gateways

TeraGrid Gateways
Last refresh: 2008-06-12T11:07:03Z

Name: Biology and Biomedicine Science Gateway

Acronym:  
Discipline: Molecular Biosciences  
PI: Daniel Reed  
Login URL: http://www.tgbiportal.org
Info URL: http://www.renci.org/

Biological and biomedical research are in transition from "small science" (i.e., research conducted by individual investigators or laboratories) to large science involving interdisciplinary teams, large-scale instrumentation, distributed data archives and multilevel computational models. Hypothesis-driven biological science, the historical modus operandi, is now being complemented by data rich, discovery-driven science. Approaches range from large-scale genome sequencing to the micro-scale assaying of the metabolic state of a single cell. Knowing which gene networks and genetic variants influence disease predisposition and progression in an organism will contribute significantly to our ability to develop effective clinical interventions. This shift to genomic medicine and intelligent hypothesis generation will lead to individual target treatments for many diseases.

Name: Chemical Informatics and Cyberinfrastructure Collaboratory

Acronym: CICC  
Discipline: Biochemistry and Molecular Structure and Function  
PI: Geoffrey Fox  
Login URL: http://www.chembiogrid.org/wiki
Info URL:

The Chemical Informatics and Cyberinfrastructure Collaboratory (CICC) uses service oriented distributed computing technology for convenient integration of distributed chemistry tools, simulations, documents and databases among themselves and with similar biological resources. We utilize open web service infrastructure and develop innovative interfaces, applications and databases for discovering the properties of small molecules.
Content in development

- HPC Local Software information
- Extended GridFTP information (Bryon Gill and SI)
- TGCP configuration information
- Core 2.0 Resource Description Repository "RDR" (Ed Hanna)
- Co/Meta-scheduling information (Warren Smith)
- SPRUCE On-Demand information (Suman Nadella)
- Science Gateway software and services (Jason Reilly, John McGee)
Soft versus Hard State

- Aggregation layer
  - MDS default is Soft state
  - TeraGrid customized Hard state

- Leaf layer
  - Determined by the provider design
Open Discussion

- Content delivery formats
  - XML
  - CSV
  - JSON
  - Perl text
- tginfo universal command line tool
- ....
Publishing New Content

1) Requirements gathering
   • Identify content
   • Information ownership
   • Information (system) sources
   • Aggregation/refresh/caching
   • Access requirements

2) Content integration
   • Is (some) content in information services
   • How is the content indexed/mapped with other content

3) Development
   • Choose existing schema and/or develop new schema
   • Use existing or develop information providers
   • Use existing or develop aggregation/refresh/caching
   • Use existing or develop access views/applications
New Content Access Views

1) Requirements gathering
   - Identify content
   - Query protocols
   - Query aggregation scope
   - Query reliability
   - Query frequency/performance
   - For users and/or software

2) Development
   - Choose existing access protocols and views
   - Develop new access views
   - Develop new access protocols
New Content Aggregation/Storage

1) Requirements gathering
   • Persistence
   • Versioning
   • Etc

2) Development
   • Extend existing aggregation/storage methods
   • Develop new aggregation/storage methods
More Information

Find out more:
http://info.teragrid.org/ (links to content and documentation)

Request content:
mailto: help@teragrid.org or navarro@mcs.anl.gov

Discuss Information Services content, requirements, and design:
E-mail list tg-cat@teragrid.org

View current Information Services content

User Portal (scheduler load & queue contents):
https://portal.teragrid.org:443/gridsphere/gridsphere?cid=resources

User Documentation (CTSS 4 kits, software, services):
http://www.teragrid.org/userinfo/software/ctss.php

Information Service Main Page:
http://info.teragrid.org/