Orchestrating caGrid services in Taverna

Wei Tan, Ravi Madduri, Kiran Keshav, Baris E. Suzek, Scott Oster, Ian Foster

wtan@mcs.anl.gov
http://www.mcs.anl.gov/~wtan

Computation Institute
University of Chicago and Argonne National Laboratory, Chicago, IL
Agenda

• Introduction of caGrid
• Why scientific workflow and Taverna?
• The caGrid plug-in in Taverna
• Case study
• Conclusion
Introduction of caGrid

• Cancer Biomedical Informatics Grid™ (caBIG)
  - Share cancer research related tools, data, applications, and services.
  - As of May 2008:
    - 62 cancer centers connected to caBIG
    - 200+ orgs participated in tools/services development

• caGrid
  - the underlying service-based grid infrastructure for caBIG.
  - Build on Globus Toolkit 4.0.
  - 80+ data or analytical services now “active”
caGrid: ecosystem and the role of workflow

There are currently 122 caBIG Participants, 81 grid services (62 data services, and 19 analytical ones).

caGrid: ecosystem and the role of workflow

To make caGrid a sustainable service ecosystem to foster “service oriented science”:
• 1) users are willing to use existing services, and
• 2) users are willing to provide new services constantly.

Scientific workflow lifecycle

- Workflow as consumer
  - Easily reuse services for complex experiments.
- Workflow as contributor
  - Workflow as “best practice” wrapped as services.
Why **scientific** workflow in caGrid?

- **Integration and collaboration**
  - to integrate a wide variety of cancer-related data and analytic services

- **Data flow oriented**
  - different modeling/execution style as opposed to the control-flow oriented business process

- **Traceability and reproducibility**
  - as evidence for scientific investigation

- **Automatic processing**
  - instead of copy & paste, switching between applications
Taverna: offers support in the lifecycle

✓ Scavenger: for customized service discovery.
✓ Feta: service annotation and discovery.

✓ Result persistence and visualization

✓ Scufl: compact modeling of data flow.
✓ Built-in processors: Soaplab, BioMart, etc.
✓ Customized processors as plug-in.

✓ Implicit iteration: handle parallel execution.

A community to share workflow

Build a caGrid plug-in in Taverna

- Taverna is not originally grid-enabled.
- It provides an plug-in framework to develop user-specific processors (caGrid services, in our case).

caGrid plug-in:
- **Scavenger**: to make Taverna aware of the caGrid services.
- **Processor**: at build-time, to configure services with Globus properties, like resource identity and security.
- **TaskWorker**: at run-time, invoke caGrid services with Globus extension.
Interactions between Taverna and caGrid plug-in

1. newScavenger()
2. query()
3. getProcessorList()
4. newProcessor()
5. wsrfConfig()
6. securityConfig()
7. invoke()
8. enforceWSRF()
9. enforceSecurity()
10. call()
Metadata-based service query

Types of query
- String based.
- Property based.
- Semantic based.

caGrid service metadata

caGrid scavenger GUI

Add Your Custom Service Query
Location (URL) of the index service: http://cagrid-index.nci.nih.gov:8080/wof/services/DefaultIndexService

Service Query Criteria:
- Research Center: Ohio State University
- Service Name: DICOMDataService
- Operation Name: PullOp

Send Service Query

WSRF and security extension

- **Web Service Resource Framework (WSRF)**
  - Provides a way to access stateful resources.
  - For example, create a counter and add value to it.
  - **Reference Properties**: identify resource instance (in SOAP header)

- **Grid Security Infrastructure (GSI)**
  - Security is important in Web-scale collaboration to protect resources across organizational boundaries.
  - **Encryption / Authentication** (identifying the caller/sender)/ **Delegation** (performing a task on behalf of a delegator) …
**Case study: microarray hierarchical clustering**

- **Microarray**
  - A high-throughput technology that measures the expression of many genes in different conditions.
  - The data from each microarray can be represented as a vector in which each element represents the expression level of a gene in a condition.

- **(Hierarchical) clustering**
  - Group similar expression vector across genes.
  - Identify genes with similar biological functions

Result Taverna workflow

- **Microarray clustering routine**
  - Querying and retrieving the microarray data of interest from a caArray data service at **Columbia University**: http://cagridnode.c2b2.columbia.edu:8080/wsrf/services/cagrid/CaArrayScrub
  - Preprocessing, or normalize the microarray data using the GenePattern analytical service at the **Broad Institute** at **MIT**: http://node255.broad.mit.edu:6060/wsrf/services/cagrid/PreprocessDatasetMAGEService
  - Running hierarchical clustering using the geWorkbench analytical service at **Columbia University**: http://cagridnode.c2b2.columbia.edu:8080/wsrf/services/cagrid/HierarchicalClusteringMage
Execution trace

Execution result as xml

1936 gene expressions visualized result

Lessons learned

- **Traditional script-based approach**
  - time-consuming and error-prone
  - cannot handle issues encountered in a Web-scale environment:
    - services are not visible or understandable to users
    - services are volatile, etc.

- **Taverna**
  - a framework covers the lifecycle of scientific workflows
  - as a container to workflows, many desirable features are available with little programming effort

- **caGrid plug-in**
  - customized service discovery, configuration and execution
Conclusion

- **caGrid**
  - Service ecosystem for cancer research collaboration

- **caGrid plug-in in Taverna**
  - Orchestrate caGrid services in an on-demand and traceable manner
  - Eases the way in which cancer scientists do routine experiment

- **Future work**
  - Enhanced support for security (SSO, authorization, etc)
  - Support for more advanced data flow patterns
    - E.g., using GridFTP for processor-processor data transfer
  - Investigate more application scenarios

- **Resources**
  - Demo: [https://webmeeting.nih.gov/p44387759/](https://webmeeting.nih.gov/p44387759/)
Conclusion

• We are gaining momentum in health informatics!
  - Tool suite provide end-to-end solution.
    - Service virtualization/discovery/composition/execution/analysis
    - Metadata/vocabulary management
    - ...
  - Recognitions from the community.
  - In collaboration with software developers, cancer research centers, hospitals, biomedical/bioinformatics data centers...
Thank you for your attention.

Contact me for collaboration and job opportunities!

wtan@mcs.anl.gov
http://www.mcs.anl.gov/~wtan