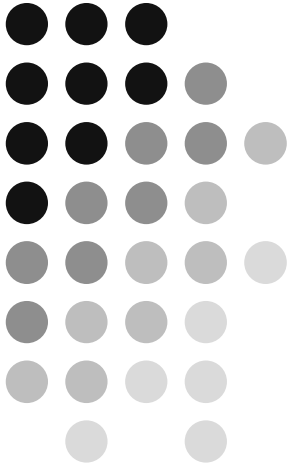
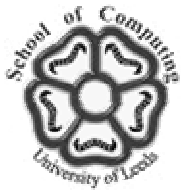
Integrating an Attack-Tolerant Information Retrieval (ATIR) Service with Taverna

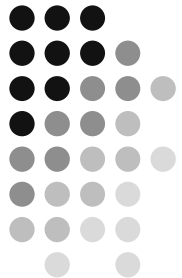
Erica Y. Yang, Jie Xu

School of Computing
University of Leeds





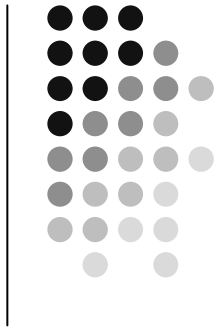
Outline of this talk



- What is ATIR?
- Why choose Taverna?
- System architecture
- Performance results
- Conclusions

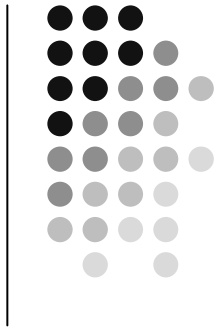
Note: This talk does not give you much details of either ATIR or Taverna

Background



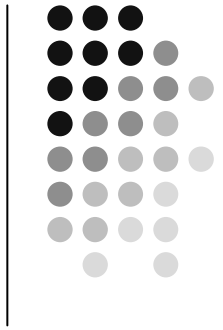
- **Realistic** Grid applications need an assured level of ***security guarantee*** about remote service providers even they are **not under the control** of users' own administrative domain (which is often the case!).
- This is particular important for commercial Grid applications that involve using **security-sensitive** data and/or processing on remote service providers.
- This requirement is **recurrent** theme for a large number of e-science/e-business applications.

Motivations



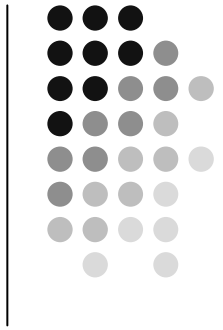
- But in an open Grid environment, such requirement is often difficult to satisfy.
  - Grid is defined by VOs which are formed based on **problem-solving scenarios** rather than a (fixed) organisation boundary.
  - Stable and consistent **trust relationships** among users and service providers simply do not exist because resources are discovered on the fly.
  - The **Grid security solutions** (e.g. GSI) that many Grid projects use are still based on the traditional security perimeter model (i.e. the system boundary is clear).

The Problem



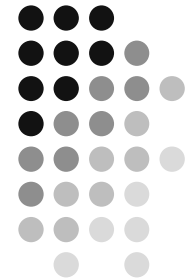
- Specifically, we address the problem of how can we securely and dependably **retrieve information** from an *untrusted* Grid environment.
- We have developed an Attack-Tolerant Information Retrieval (ATIR) technique to solve the above problem through:
  - Protect the privacy of *users* against untrusted Grid nodes (or Grid service providers)
  - Effectively detect or mask any job tampering against both intentional and unintentional faults, and/or
  - Obtain correct results (information) even in the presence of malicious attacks
- We aim to demonstrate the application of ATIR through a realistic e-Science application – Taverna.

Application Domains



- Privacy-critical applications
  - (Bio-) Medical information services
  - Computationally-intensive drug discovery by pharmaceutical companies
  - Credential services in the Grid, e.g., online Certification Authority (CA)
  - Financial information services
- Information sharing among untrusted parties
  - Robust logging facilities for the Grid
  - Dynamic collaboration over the Grid
  - Ad-hoc collaboration
- Global and remote computation
  - Mobile agent applications

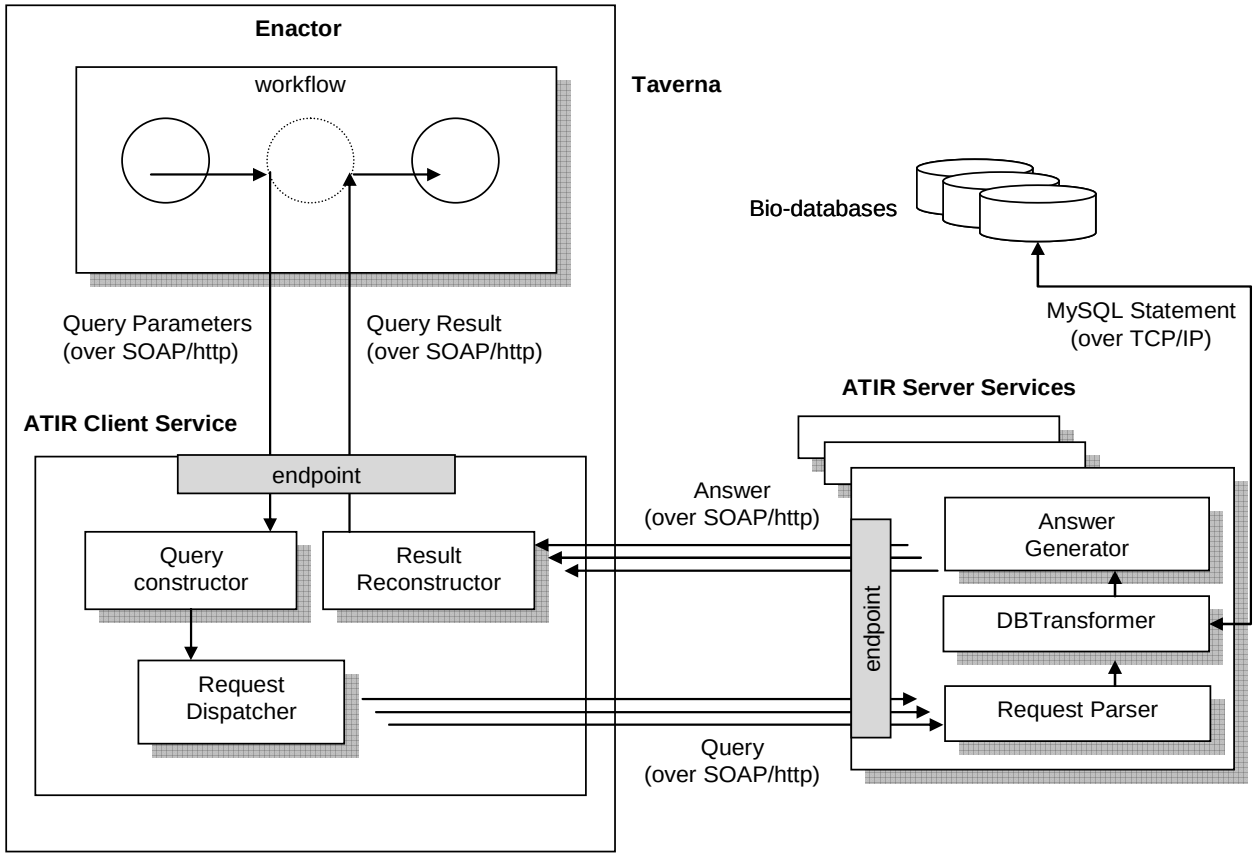
Integrating ATIR with Taverna



- Taverna, as a component of the myGrid project, is a *highly successful workflow tool* to “facilitate easy use of workflow and distributed compute technology within the eScience community.” It supports “typical bioinformatics analysis tasks”.

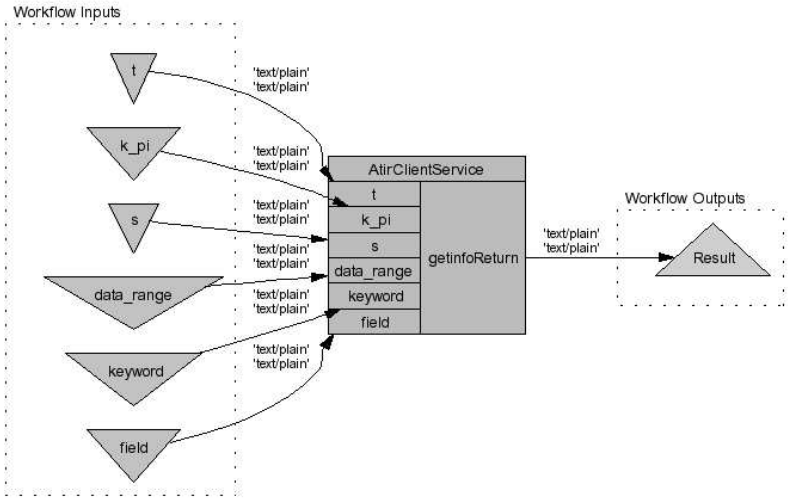
- Why Taverna?
  - Its popularity and its importance to UK e-Science community
  - Importance of its application domain
  - The lack of security provision in its current release
  - Its involvement of large-scale genome databases
  - Support of information retrieval operations

System Architecture



This figure shows the system architecture of the security-enhanced Taverna system.

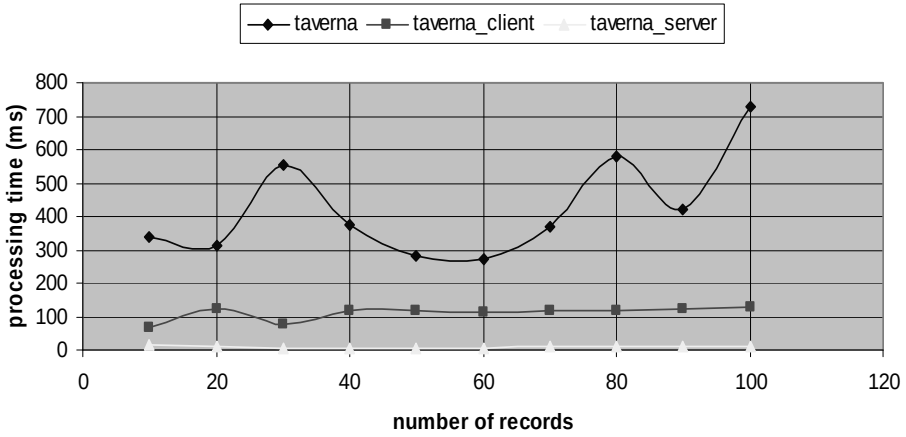
A Simple Workflow



- It takes six query parameters as workflow inputs, they are: number of curious servers that the system can tolerate (*t*), minimum number of servers (*k_pi*), the size of a result set (*s*), valid range of data (*data_range*), intended data item (*keyword*), and intended record columns (*field*).

- The size of a result set is a parameter that is chosen by the user (or the previous workflow processor) to determine the level of privacy protection that is required.

Performance Evaluation (I)



• The proportions of the ATIR performance overhead is relatively insignificant to the total processing time of Taverna.

• Consistently across all measurements, ATIR client time is less than 40% of the total processing time.

• ATIR server time trivial (<5%) while it is compared with the total processing time.

Note:
1. The service was invoked over the Internet.
2. Measurements are in milliseconds.
3. Performance results are also available for the C and Java ATIR implementations.

Performance Evaluation (II)

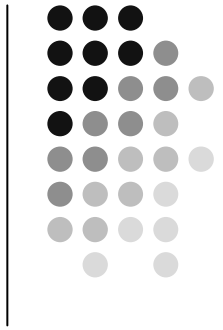
s	Tam	Tt	Ttp/Tt
10	157.42	338.9	54%
20	92.8	310.9	70%
30	126.5	551.6	77%
40	128.2	371.8	66%
50	127	282.8	55%
60	130.9	270.3	52%
70	126	367.3	66%
80	130.3	578.3	77%
90	153.5	421.9	64%
100	173.1	726.5	76%

s: number of records involved in the server side computation
T_{am}: execution time of ATIR service
T_t: total processing time of the integrated system
T_{tp}: time consumed by the Taverna enactor for parsing WSDL files and generating XML files.

$$T_t = T_{am} + T_{tp}$$

• More than 50% of the total processing time of the integrated Taverna is consumed by the enactor for WSDL/XML parsing and generating.

Conclusions



- Complemented by the performance evaluation, an application of ATIR has been successfully demonstrated on a realistic e-Science scenario.
- Integration is straightfoward (for Web services ready applications, integrating the ATIR technique is simple).
- For the experiments we performed, the overhead of ATIR is less significant compared with the total processing time of the security-enhanced Taverna system.

Demo: available at the White Rose Grid Stand.