A Service for Biological Database Replication and Update

Vincent Breton on behalf of Jean Salzemann and Nicolas Jacq
LPC IN2P3/CNRS

Institut National de Physique Nucléaire et de Physique des Particules
Information Society and Media

Introduction

- Data integration is a key challenge to bioinformatics
- Biological data bases contain all the data needed by biologists for their analyses
  - Biologists can’t do research without proper access to them
  - All molecular biologists need to access at least one database for their research
- These databases have several properties
  - Most of them have different data models
  - Most of them have different semantics
  - They are ALL growing very quickly
- Goal: provide transparent access to relevant versions of all needed biological databases
  - To enable biological analysis
  - To enable workflows

Grid enabled biological data integration

- The challenges
  - The databases keep growing very quickly AND the biologists need to access the most updated version
    - The biologists may also need to access an old version to check previous results
  - Databases need to be indexed
    - Computing intensive task
  - The data models keep evolving
- The grid added value
  - Grids can provide tools to replicate files automatically
  - Grids can provide computing resources for database indexing
  - Web services can be used to present standardized interfaces to databases

Goals of the service

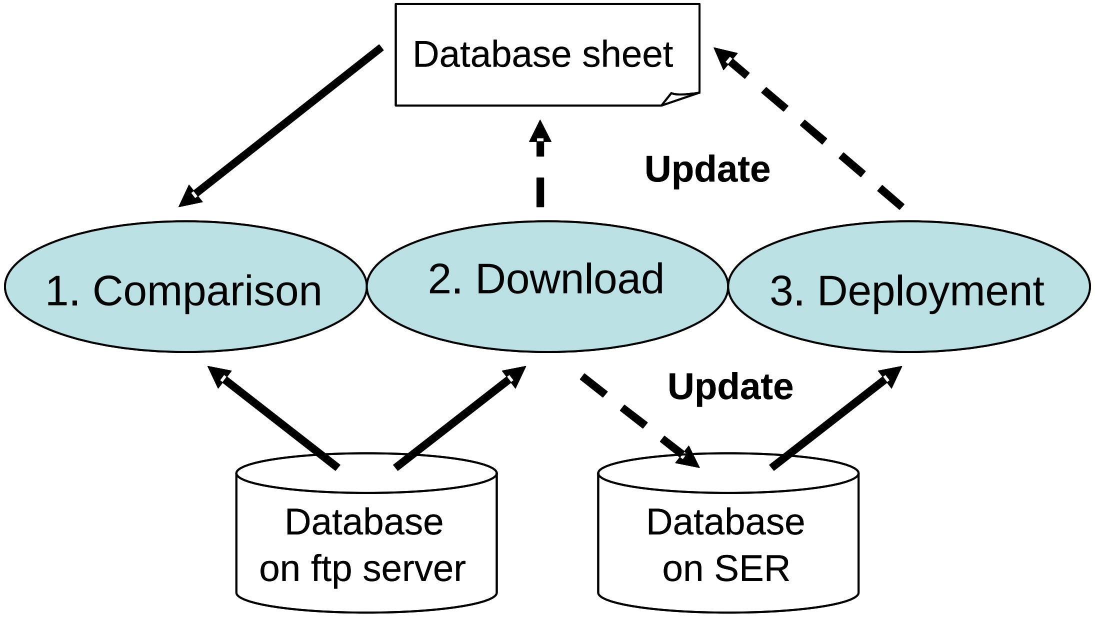
- to provide the grid users the most up to date version of any biological database
- to do it transparently
- To do it without disturbing the running jobs

RUGBI, a grid for bioinformatics in France

- Project funded by the french ministry of research (2002-2005)
- Regional grid in Rhône Alpes and Auvergne
  - Limited number of sites
  - Open and heterogeneous grid: public and private research, interoperability
- SMEs (Biopôle de Clermont Limagne)
  - Security and confidentiality
  - Transparency and easy use, bioinformatics services
  - Large storage and computing resources
  - Additional services : mutualisation , collaboration, hosting
- Pre-competitive
  - Exploitation, administration and monitoring facilities
  - Quality of service, business model

Service developed within Rugbi

- Biologists are using, most of the time flat files databases available on ftp repositories.
- The service developed is an applicative service, integrable in a grid environment, which performs automatically regular updates and propagate them through the grid
  - Management of jobs using old version of databases
- The service does not keep previous versions of the databases

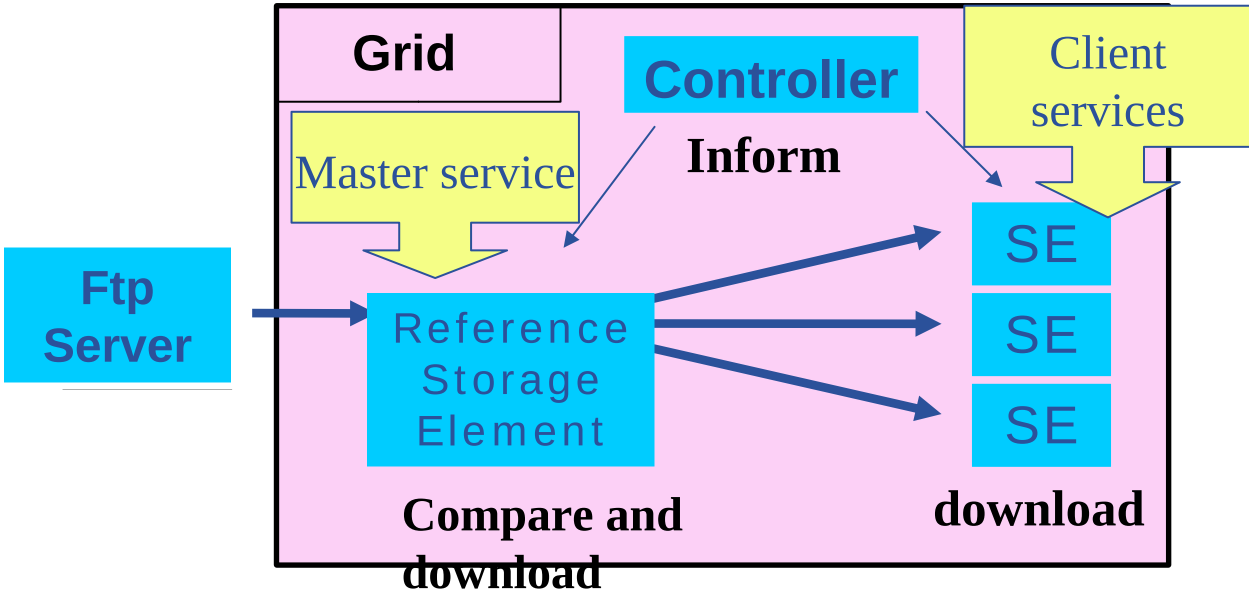


Database description on the grid uses XML sheet

Service concept

- Master Service:
  - Get the information from the information system (Controller)
  - Compare the states of the databases
  - Download the differences
  - Notify the clients
- Client Service:
  - Get the information from the information system
  - Download the differences

- Implemented in java as web Services and tcp socket.
- Compatible with Axis, Globus Toolkit 3, Globus Toolkit 4.



Grid

Controller

Client services

Master service

Inform

Ftp Server

Reference Storage Element

SE

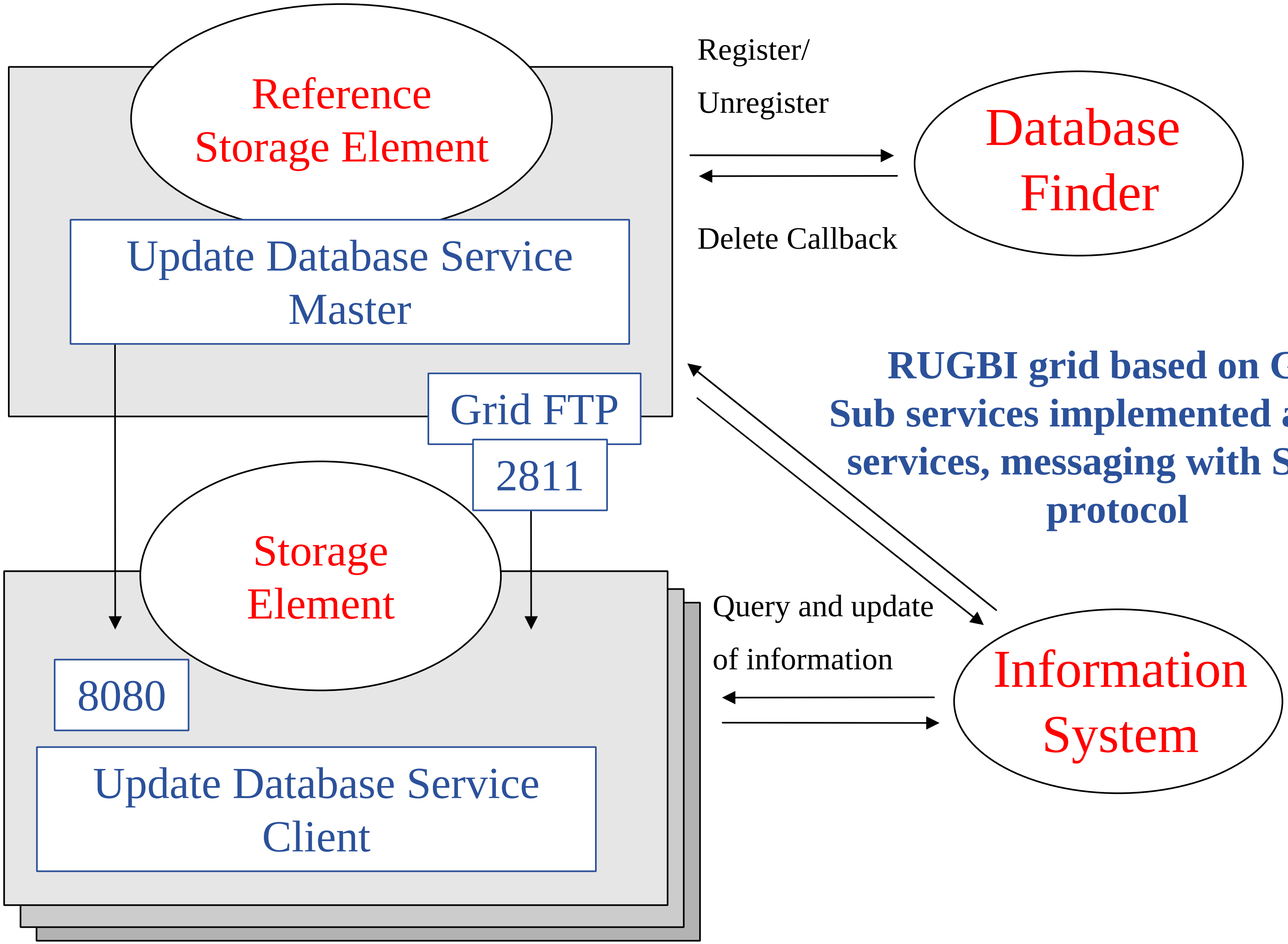
SE

SE

Compare and download

download

General Architecture in Rugbi Grid



RUGBI grid based on GT4
Sub services implemented as web services, messaging with SOAP protocol

Main Steps of the process

1. The SER updates its repository and notifies the clients (Comparison + download)

2. The SE gets the notification and download the updates with GridFTP.

3.The SER ask for a REGISTER of the new database and an UNREGISTER of the old version.

4. The SE notifies the success of the deployment to the SER

5. The SER is waiting for a deletion notification of the old version, when it is received, it deletes the old database and propagates this notification through the grid.

The data

- The databases were selected based on end users requirements (Biotech SME’s, public labs)
  - Swissprot, 700 MB
  - Trembl, 2.4 GB
  - Pdb, 2.9 GB
  - Kegg, 13 GB
  - **Embl, 476 GB , 180 GB (release, without annotations)**

- Possibility to add new databases.

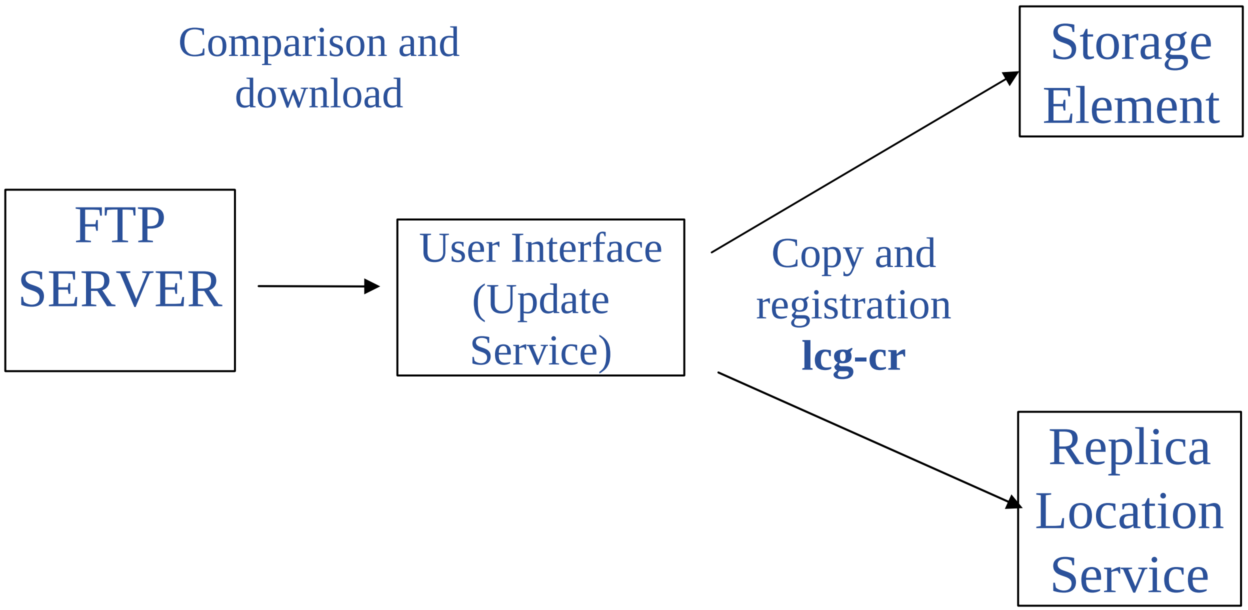
- The databases are described as dynamical XML sheets, containing all the necessary information to make each step of the process.

Pre-deployment XML example

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE database SYSTEM "db.dtd">
<database id="21" name="EMBL">
  <characteristics category="DNA" checked="Unknown" creation_date="20/10/04"  description="EMBL"
     update_date="15/09/05" version="0">
      <copyright category="free" user_type="all" weburl="http://www.ebi.ac.uk"/>
  </characteristics>
  <deployment type="flat_file">
      <install required_architecture="none" required_dbms="none"
          required_mb_space="200000" required_platform="none">
          <download dbroot="/pub/databases/embl/" protocol="ftp" type="original"
     url="ftp://ftp.ebi.ac.uk/pub/databases/embl/">
              <target name="/pub/databases/embl/release/" path_depth=”0” />
          </download>
      </install>
      <structure/>
  </deployment>
  <use ontology="yes"/>
</database>
```

Service deployment

- Deployment on the Rugbi GRID
  - eight sites in Clermont-Ferrand, Lyon and Grenoble
  - databases deployed and updated regularly: SWISSPROT (700 MB), TREMBL (2.4 GB), EMBL (release without annotations: 180 GB), KEGG (13 GB), PDB (2.9 GB), NCI (900 MB)
- Deployment on Auvergrid and EGEE
  - Requires interfacing the service with EGEE services (information system, data management)



Perspectives (I/II): Embrace

- Embrace is a Network of Excellence funded for 4 years by DG-RTD since February 2005
  - 17 partners, coordinator: EBI
- Embrace aims at building a « knowledge grid » allowing integrated exploitation of biological data
- Year 1 was dedicated to understanding the environment
  - Identification and description of test cases including database replication
  - Evaluation of the existing infrastructures
  - Creation of a virtual organization on EGEE as test bed
- Year 2 will be dedicated to initiate the building of the Embrace Grid
  - Technology recommendation (web services)
  - Implementation of test cases
  - Analysis of biological grand challenges to be deployed on Embrace grid

Perspectives (II/II): BioinfoGRID

- BioinfoGRID is about promoting Bioinformatics Grid applications for life science
  - Genomics
  - Proteomics
  - Transcriptomics
  - Molecular Dynamics
- WP4 is dedicated to studying distribution of biological databases on the grid
  - Collaborative work with Embrace and EGEE

Conclusion

- Grids open new opportunities for integration of biological data
- Development of a database replication service within the framework of the French research project RUGBI
  - Built on web services
  - Adaptable to existing grid middlewares
  - Only the last version of each database available on the grid
- Perspectives
  - On-going activity within BioinfoGRID, Embrace and EGEE-II European projects
  - Joint workshop on Grid data replication, consistency and requirements in Pisa May 26, 2006