Lightweight Solution for Protein Annotation

UK e-Science All Hands Meeting, September 2005

Shikta Das

Bioinformatics Research Assistant

London e-Science Centre

Imperial College London

Contents

- The e-Protein project
  - objectives of the e-Protein project
- Why structure-based proteome annotation ?
  - The annotation pipeline at IC
  - 3D-GENOMICS database
- Challenges
- Implementation
- Conclusions
- Acknowledgements

e-Protein: www.e-protein.org

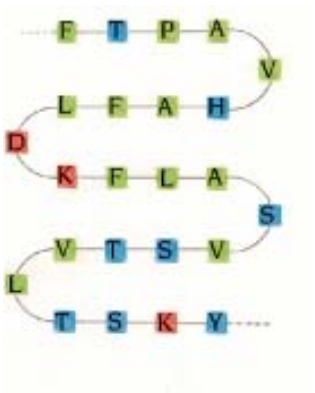
Mission statement – "*To provide a fully automated distributed pipeline for large-scale structural and functional annotation of all major proteomes via the use of cutting edge computer GRID technologies*."

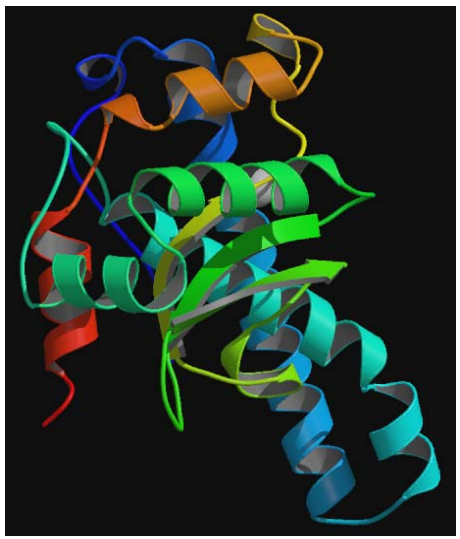
- Funded by Biotechnology and Biological Sciences Research Council/ Department of Trade and Industry (BBSRC/DTI) through their e-Science program
- Distributed Pipeline for structure-based proteome annotation using Grid Technology at three sites
- Three sites involved – Imperial College London (IC), University College London (UCL), European Bioinformatics Institute (EBI)
- Annotation pipeline utilises homology and fold recognition methods

Objectives of the Project

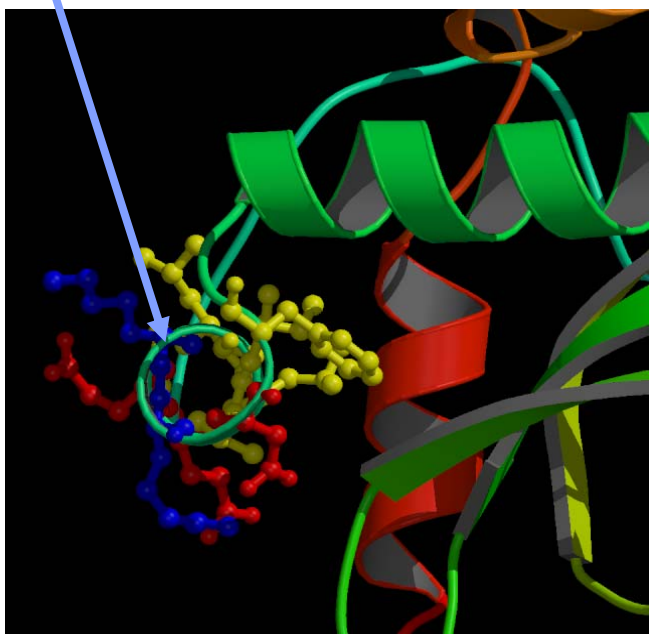
- Establish local databases and disseminate to biological community via distributed annotation system (DAS)
- Comparison of alternative approaches for annotation
- Share computing power transparently between sites using GLOBUS
- Use of robust Grid technologies
- Prototype for a national distributed proteome annotation Grid

Why Structure-based Annotation?





Enzyme active site

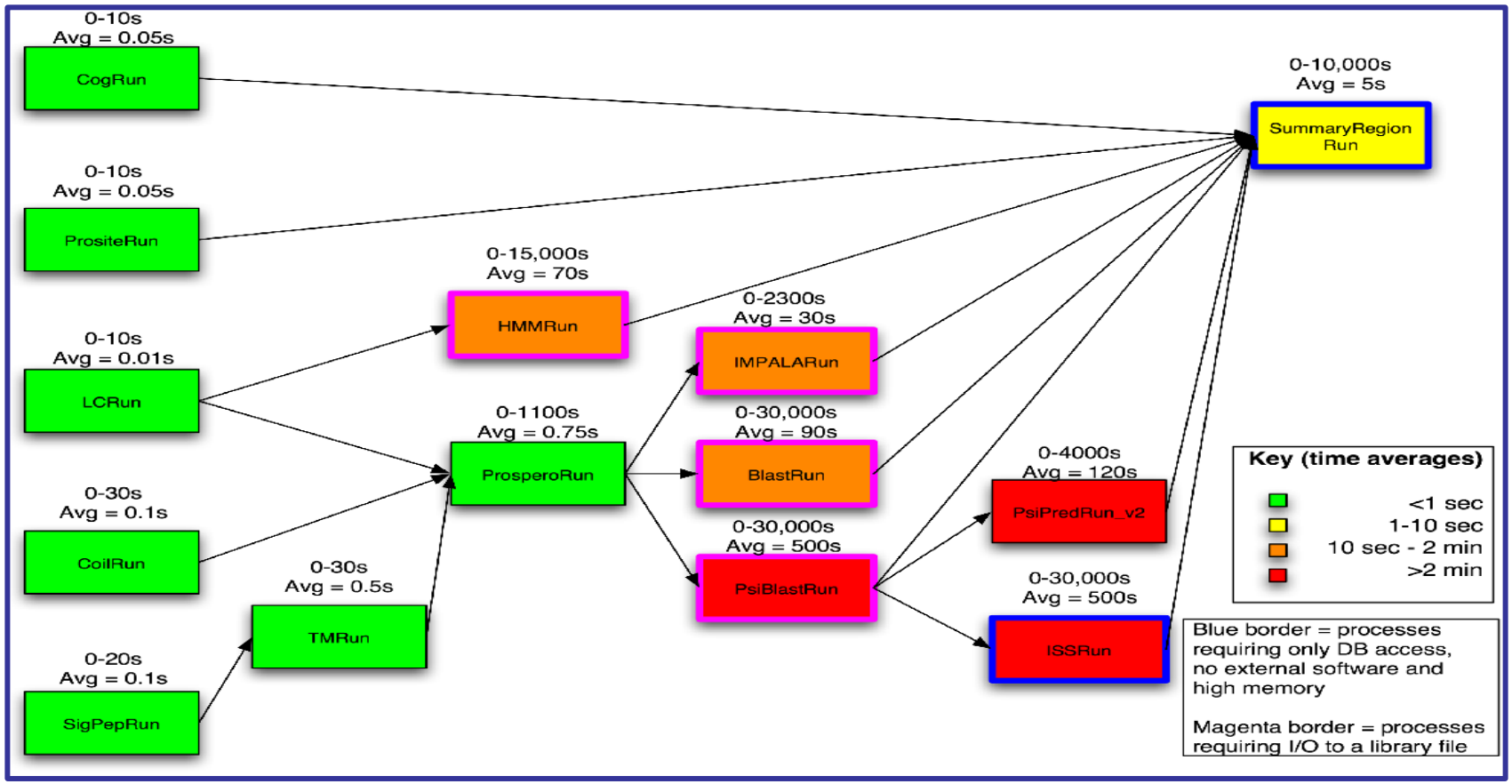


Protein Sequence

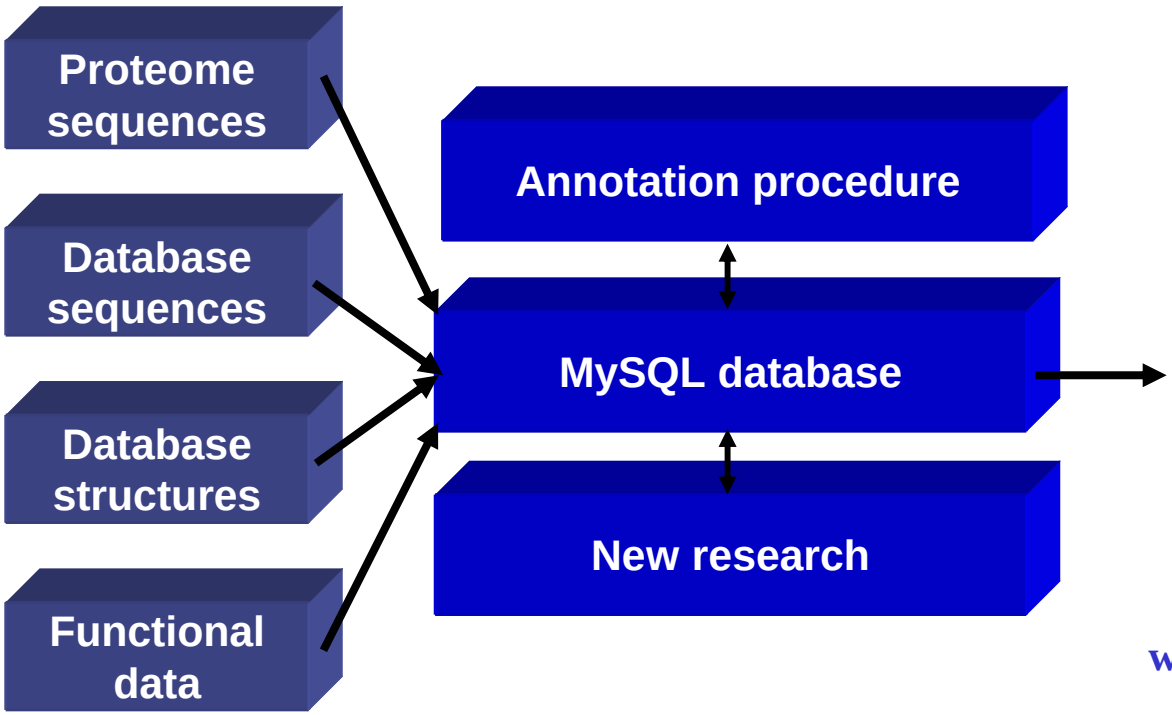
Protein Structure

Residues involved in enzyme catalysis

The Annotation Pipeline at IC



3D-GENOMICS Database of annotation



www.sbg.bio.ic.ac.uk/3dgenomics

Dr Victor Lesk, Imperial College London

Challenges

Version 1.4 of ICENI:

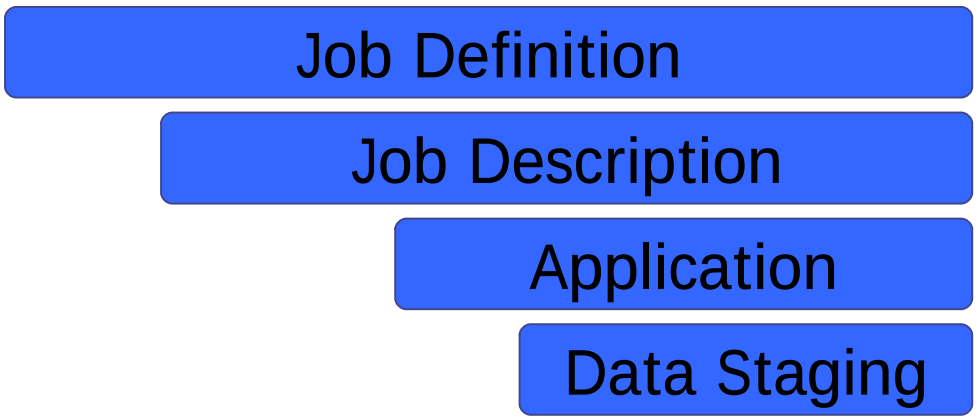
- Overburdened due to software decay
- Cumbersome to install
- Problems executing jobs on remote resources due to Java Jini firewall issues

Implementation of ICENI II

- "GridSAM" is one of the services featured in ICENI II
- Funded by Open Middleware Infrastructure Institute (OMII)
- Submission and monitoring of jobs
- Transparently submits jobs to Distributed Resource Management (DRM) systems such as *Condor* and *SGE*
- Deployable on any Java Servlet compliant container

Job Submission Descriptive Language (JSDL)

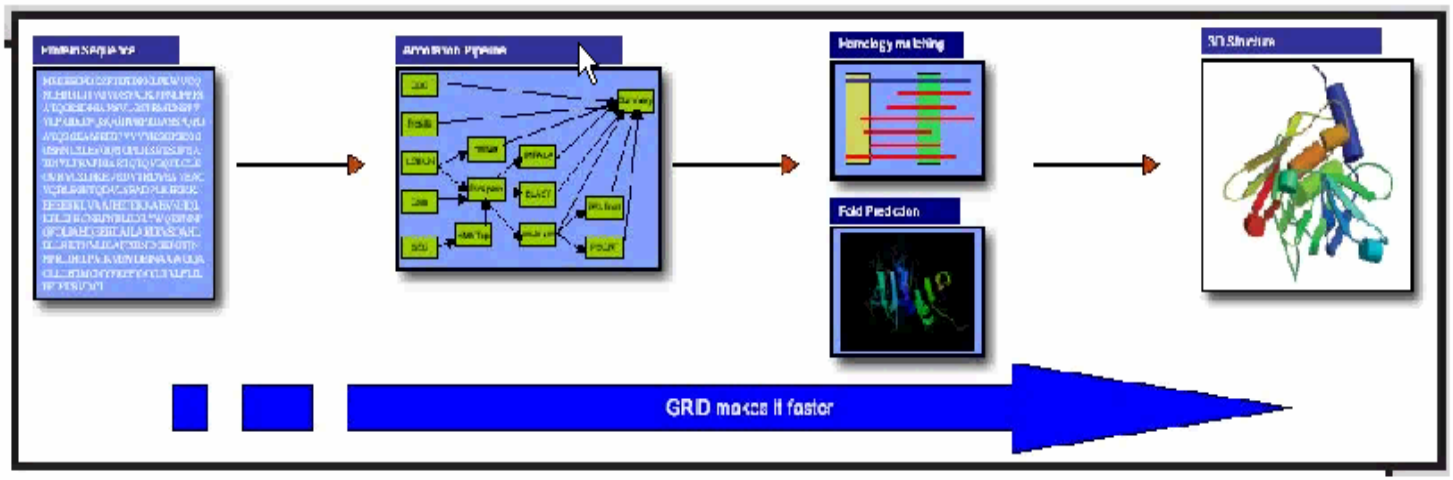
- XML template language for describing core aspects of a job



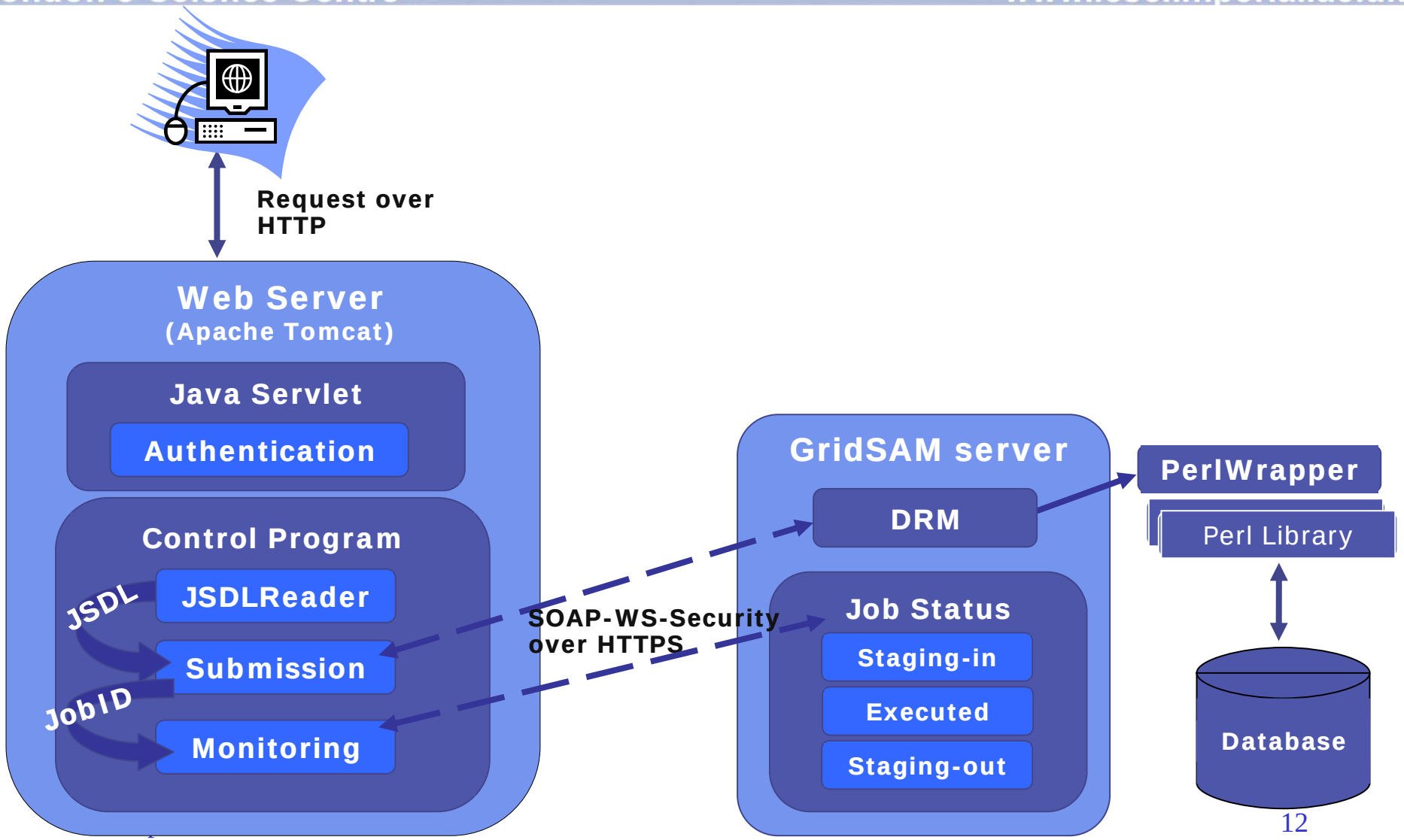
Home Page
Projects
Supported Activities
Resources
Services
News and Events
Publications
ICENI- Grid Middleware
Articles and Links
Current Vacancies
Contacts

3D-Annotate

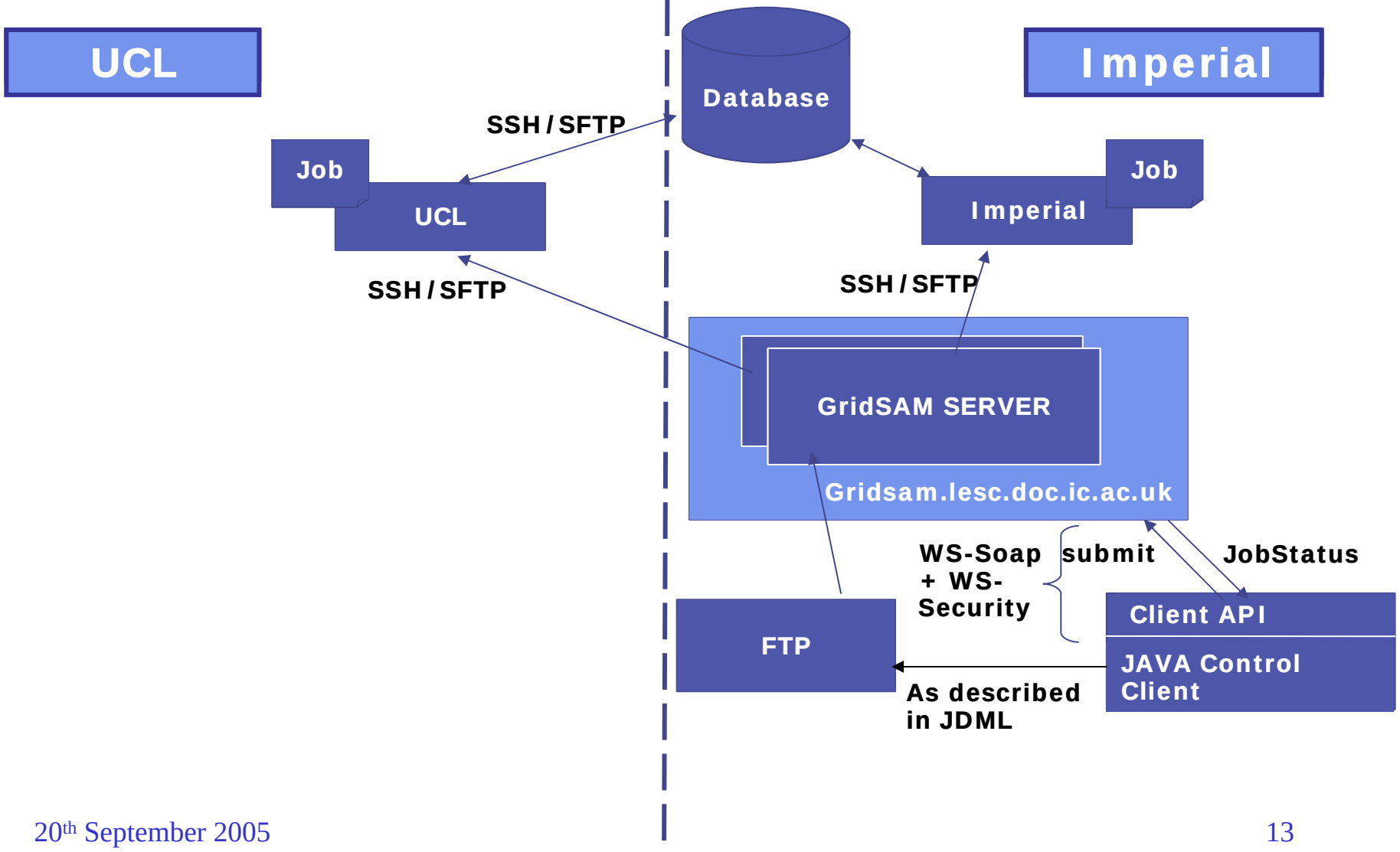
The e-Protein project is using emerging Grid technologies to combine heterogeneous resources at multiple sites (Imperial College London - IC, European Bioinformatics Institute - EBI, University College London - UCL) collaborating in the execution of these proteome annotation pipelines. The pipeline used within the project is utilising homology and fold recognition methods to assign structures to the proteomes and generate three-dimensional models. The London e-Science Centre is developing ICENI a service-oriented middleware framework which is used extensively within the e-Protein project to capture the workflow of this pipeline, and map it to resources on the Grid.



Login Page


Request over HTTP
Web Server
(Apache Tomcat)
Java Servlet
Authentication
Control Program
JSDL
JSDLReader
Submission
JobID
Monitoring
SOAP-WS-Security over HTTPS
GridSAM server
DRM
Job Status
Staging-in
Executed
Staging-out
PerlWrapper
Perl Library
Database

Grid Sharing



Conclusion

- GridSAM is a “Lightweight” solution:
  - It provides pluggable submission pipeline to connect to variety of DRMs
  - It uses HTTPS transport security and WS-Security framework for authentication and authorising users
  - Fault-tolerance by long-term persistence of job states
- Future work:
  - Provide regular updates to the database
  - Investigate issues concerning network security
  - Include error check at submission level
  - Utilise other modules of ICENI II

Acknowledgements


Prof John Darlington


Dr Andrew S McGough


William Lee


Dr Keiran Fleming


Jeremy Cohen


Oliver Jevons

- Prof M Sternberg
- Prof D Jones
- Prof C Orengo
- Prof J Thornton