Structural Biology in the context of EGEE

Germán Carrera, David García, Jr. Valverde, Jakub T. Moscicki, Adrian Murau, Jose-María Carazo

10 May 2007, EGEE User Forum #2 Manchester

www.eu-egee.org

Information Society and Media

Biological Problem

- **3D Structure refinement**
  - X-Ray, NMR, 3D-EM,...
- **Structure prediction**
  - Homology modelling, threading, motif detection...
- **Structure modelling**
  - Molecular mechanics, Molecular dynamics, Quantum mechanics...
- **Structure analysis**

Biological Problem

- **The Cryo-electron microscopy permits structural characterization of macromolecular assemblies in distinct functional states.**

The aim is to obtain a computer reconstruction of the specimen (Ribosome, Antigen ...) but:

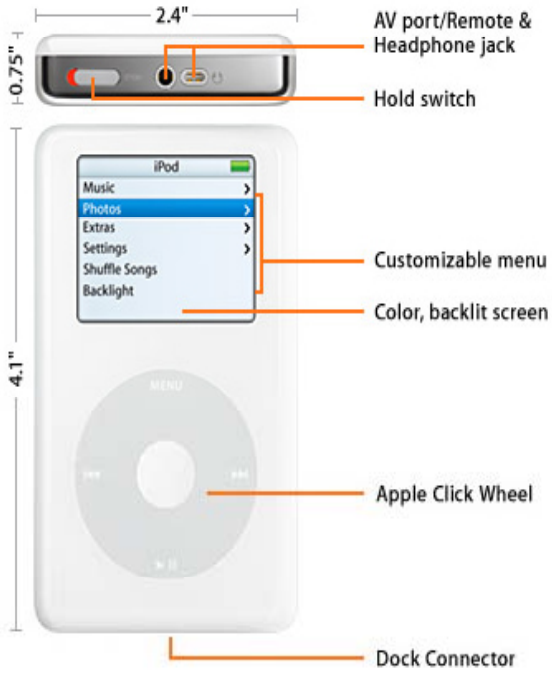
- **the images are extremely noisy (SNR<1/10!)**
- **the projection directions of the experimental images are not known**

Biological Problem

- Many samples show structural heterogeneity,(e.g when more than one conformational state of the complex is present)



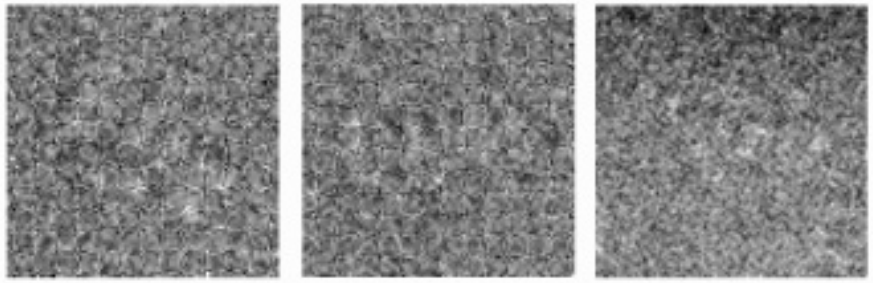


- **IPod is a trademark of Apple Corp**

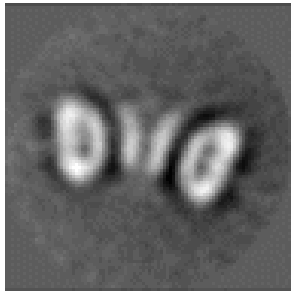
Biological Problem

- **Fight the noise by averaging**



average over

2,000 copies



- **BUT, this requires:**
  - alignment: determine the unknown orientations
  - classification: separate distinct conformations

Maximum Likelihood

- **Optimize the probability of observing the data, given the model (i.e. *the likelihood*)**
- **These probabilities are based on:**
  - a statistical model of the experimental noise
  - a statistical model of the orientations & classes
- **Mathematics: optimizing the likelihood yields models with less bias than any other method (Rice, 1995)**

First Applications

- **Selected by their user impact (Scheres et al., 2007, Nature Methods. Selected by the Faculty of 1000 as a "Must Read" article)**
- **MLalign3D**
  - The combination of images in a 3D reconstruction requires:
    - That they represent projections of identical 3D objects.
    - That their relative orientations be known.
  - MLalign3D combines the tasks of
    - Classifying the images in homogeneous groups.
    - Aligning the images to obtain the best orientation.
- **MLalign2D Is a similar to the 3D alignment (simpler)**

Computational Aspects

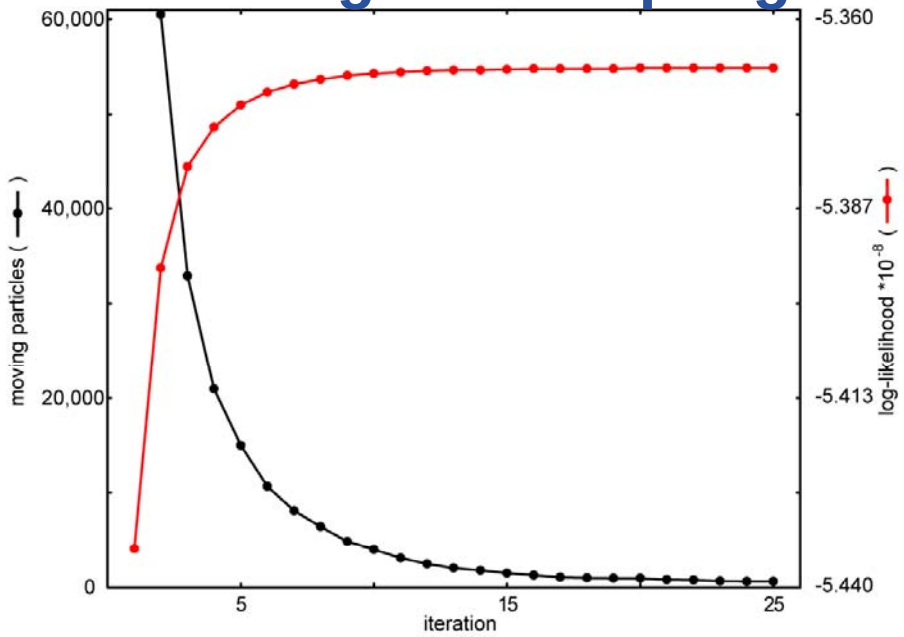
- **MLalign3D and 2D algorithms demands**
  - Large memory requirements
    - Typically we distribute an amount of 100.000 images with sizes that can be [150x150] in the 3D case.
  - High computational resources
  - Align programs are characterized by a relatively large set of free parameters.
  - Maximum-likelihood method requires long time processing.

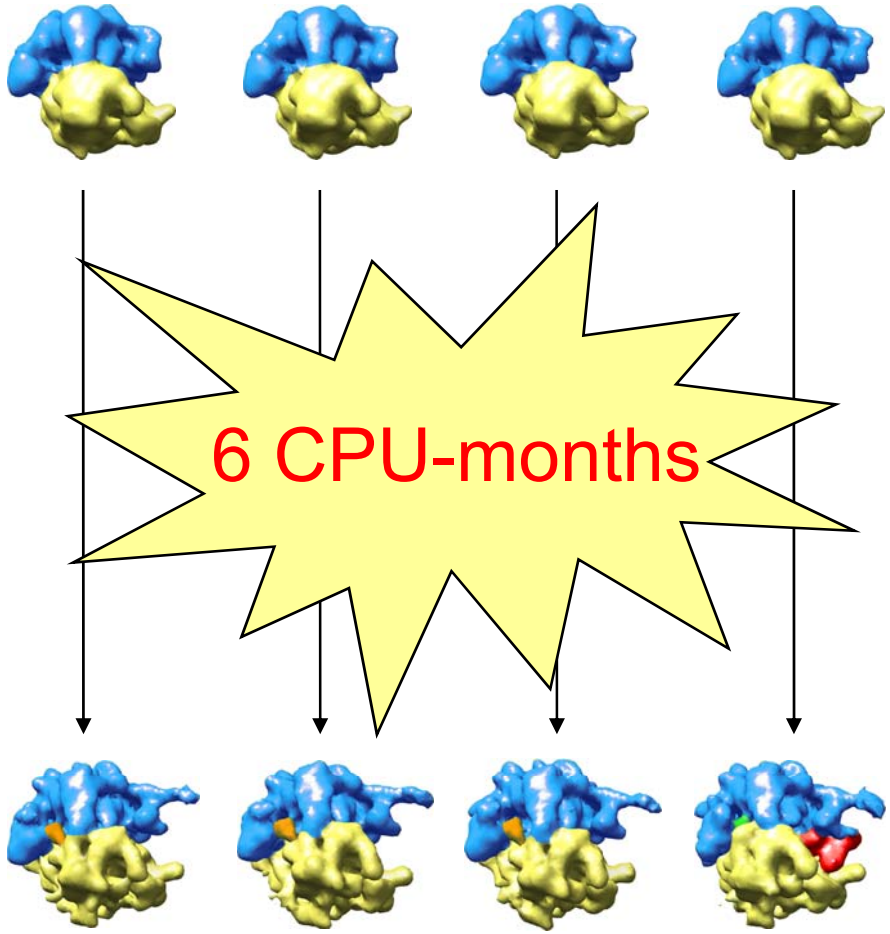
But it is expensive!

- **A typical 2D run:**
  - 10,000 images, 10 references: ~CPU weeks
  - A typical 3D run:
  - 50,000 images, 4 references: ~CPU months

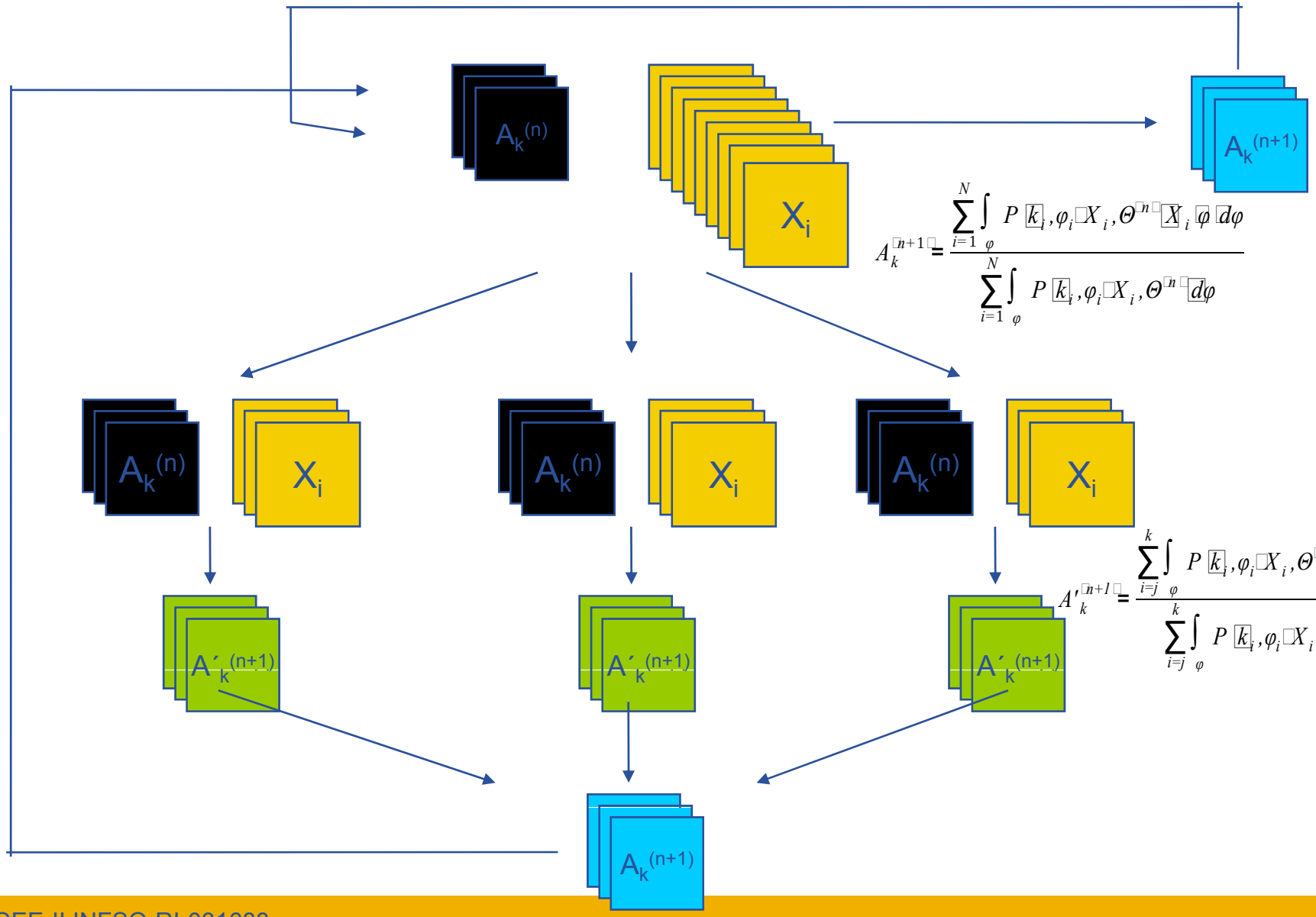
ML3D-classification

- 4 references
- 91,114 particles
- 64x64 pix (6.2Å/pix)
- 25 iterations
- 10° angular sampling





Paralellization strategy

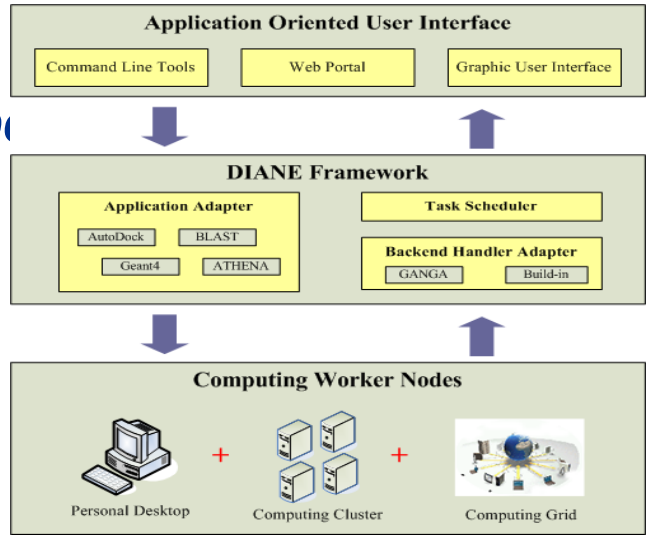


$$A_k^{n+1} = \frac{\sum_{i=1}^{N} \int_{\varphi} P(k_i, \varphi_i | X_i, \Theta^n) X_i \varphi \, d\varphi}{\sum_{i=1}^{N} \int_{\varphi} P(k_i, \varphi_i | X_i, \Theta^n) \, d\varphi}$$

$$A'_k{}^{n+1} = \frac{\sum_{i=j}^{k} \int_{\varphi} P(k_i, \varphi_i | X_i, \Theta^n) X_i \varphi \, d\varphi}{\sum_{i=j}^{k} \int_{\varphi} P(k_i, \varphi_i | X_i, \Theta^n) \, d\varphi}$$

Porting Solution

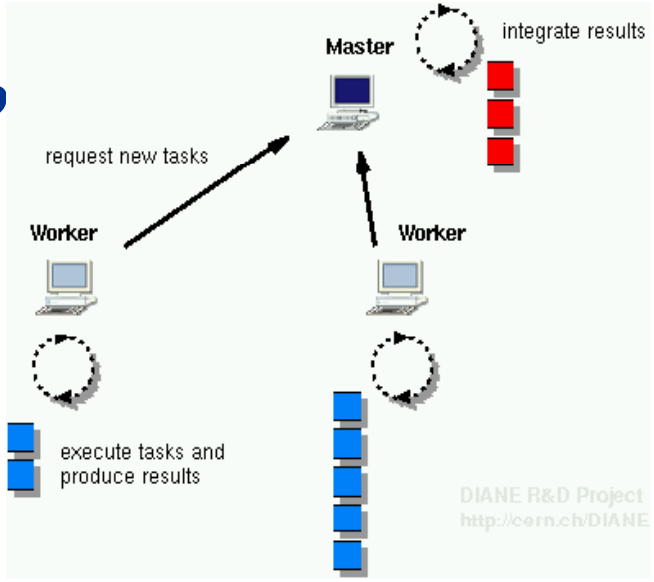
- *DIANE framework*
  - ***DIANE is a lightweight distributed framework for parallel scientific applications. It assumes that a job may b split into a number of independent tasks which is a typical case in many scientific applications***



- *DIANE*
  - ***DIANE is written in the python language.***
  - ***You can deploy your own module (in python, C/C++ ...) for a specific application.***
  -

Porting Solution

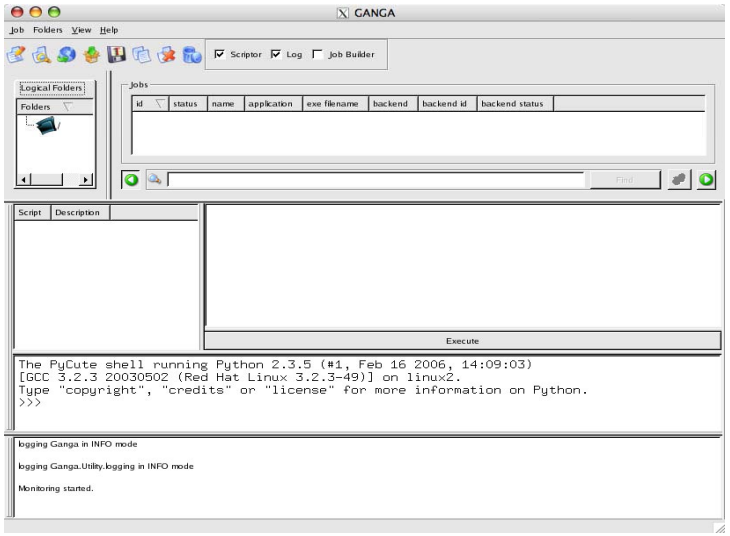
- *Why DIANE?*
  - ***We use DIANE to improve and manage the execution of MLalign on splitting mode.***
  - ***DIANE is based on a pull model useful to obtain better performance.***



- *DIANE operation*
  - ***Master-Worker Workflow Model***
  - ***DIANE is based on pull model***
    - ***workers ask for tasks to the master***
    - ***Master decides how to assign tasks to workers***
    - ***and user may optimize this process for a particular application.***

Porting Solution

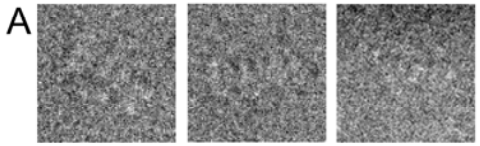
- **GANGA**
  - GANGA is a front end for job definition and management of analysis jobs to run locally and in an distributed environment. It helps
    - in the creation and configuration of analysis job
    - Submission of the jobs and finally monitors job status and take care of saving any output.

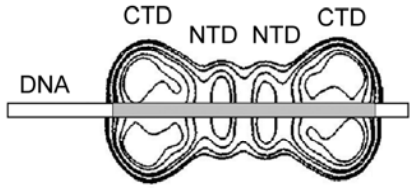


- Ganga and DIANE are developed and supported by ARDA project at CERN (EGEE NA4)
- Scalability of DIANE (security disabled): up to 400 workers nodes simultaneously, server handling tasks at 100 Hz
- Security may be enabled and is based on SSL and Grid Certificates
- The Ganga interface is the data analysis solution for Atlas and LHCb

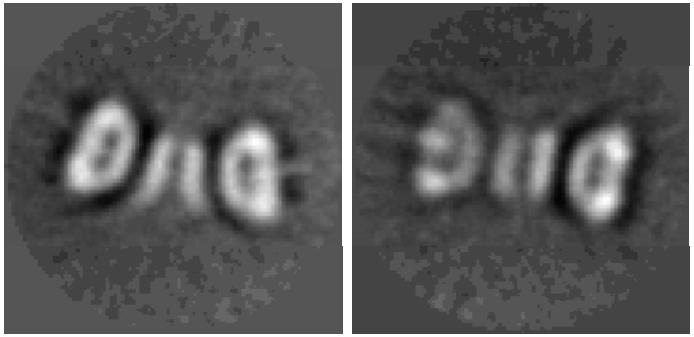
Demo's case:

- **The machine that starts replication of a virus (SV40) that causes cancer…**
- **1,670 projections of a complex with an asymmetric DNA-probe**





- **To limit computation, we will use 2 references, to reproduce:**



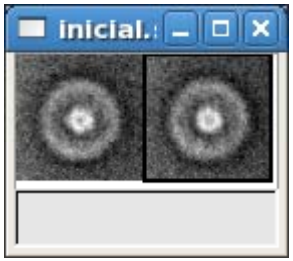
Results

- **The complete convergence of this set of images has been reached in 28 iterations.**
  - The initial request for DIANE has been of 50 worker nodes for this case.
  - The duration of each iteration has rounded 2~3 minutes using different sites.
  - A better performance has been obtained sending the images only in the initial iteration

Results

- Evolution process

Initial references



Iteration 1

Times of some iterations

17:27 17:34

17:34 17:37

17:37 17:40

17:40 17:42

17:43 17:45

17:46 17:48

…….

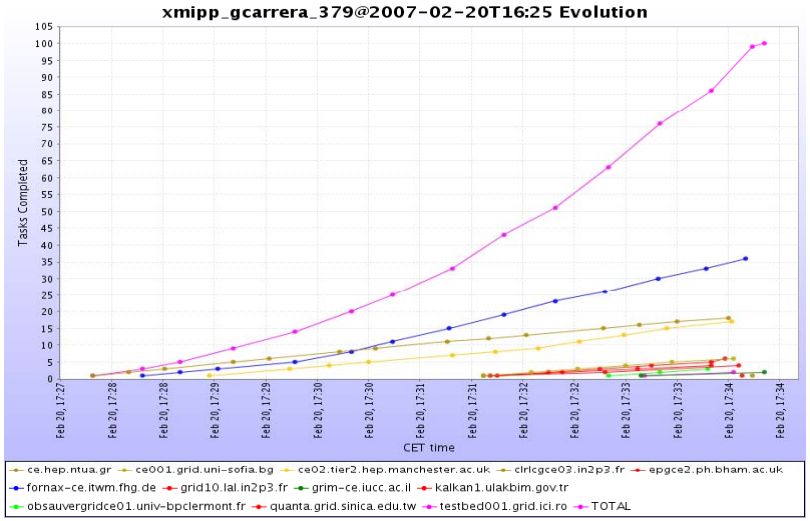
17:55 17:57

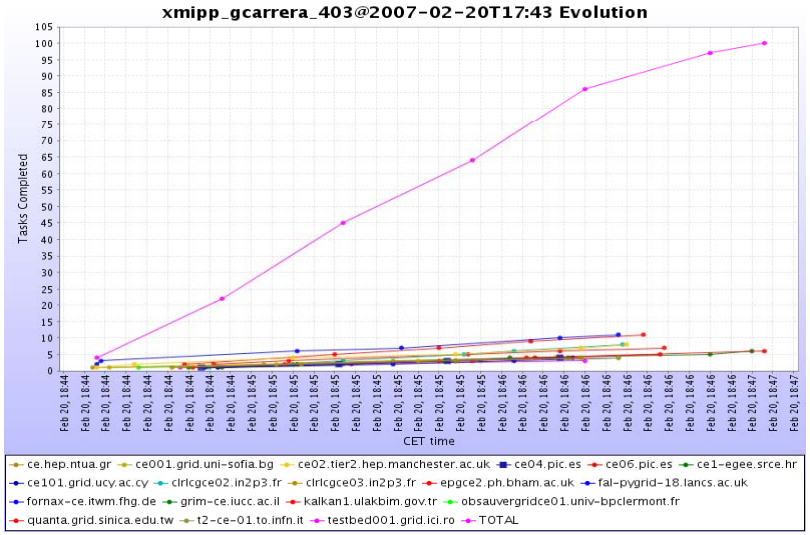
17:58 18:00

18:01 18:03

18:04 18:06

Iteration 24





Structure prediction

- **YaMI (Yet another Modeller Interface)**
  - We adapting our previous web interface to Modeller to launch jobs on the Grid
    - Make easy for end users to use Modeller
  - Modeller is a highly popular tool for doing
    - Molecular model construction by homology
    - Molecular modelling

Molecular Modelling

- **AMBER**
  - Is one of the most popular molecular modelling tools (molecular dynamics, molecular mechanics, QM/MM)
  - Large problems require MPI over large clusters
- **TINKER**
  - A popular, easy to use molecular modelling package
  - Includes programs for most common modelling tasks
  - Generates large trajectory files

Molecular interactions

- **GROCK is a tool to explore molecular interactions**
  - Allows docking a molecule against a database of structures
  - Predict protein-protein, protein-ligand, ligand-protein interactions
  - Launches thousands of jobs
    - Some jobs (e.g. FTDOCK) may take days to run
  - Its latest release adds
    - Enhanced support for error recovery
    - Support for 3D-Dock (and additional databases)

Course Experience

- **Last Friday 23rd of February three European Union projects converged at the National Center for Biotechnology, CNB - CSIC, in Madrid**
  - Network of Excellence on “3D Electron Microscopy”
  - EMBRACE
  - EGEE

Course Experience

- **Some of the subjects that were treated in the Embrace part of this course were the following ones.**

- **Workflows in bioinformatics**
- **Web Services**
- **BioMOBY**
- **DAS - 3DEM**
- **EGEE Grid overview**
- **2D/3D aligment introduction**
- **A joint practical session between the assistants of the two disciplines**

Course wrapup

- **Course conclusions:**
  - The students were satisfied (contents and depth)
  - They threw in lack more practical cases
  - Each student obtained a certificate in BIOMED vo and an account in our UI to continue working
    - To catch new developers

Requests

- A better support for longer jobs, In our case we could need to retain resources during weeks. (eg: long queues > 72 hours).
- It's necessary to improve the communication between the middleware and the developer through a more detailed and reliable output of the lcg/edg/glite commands.
- A better communication and collaboration between all the EGEE project participants with the aim to obtain a Grid well suited for large scale scientific problems.

Requests

- **Data Management improvements:**
  - Some POSIX solution like ELFI (user file system) with encryption methods: a real shared *filesystem* between the different storage elements. (e.g. LFC as a POSIX filesystem, PVFS, GFS) Is this possible?
  - User data separation, virtual workspaces...
- **Better MPI support, I recognize a great effort has been done (by Cal Loomis & Stephen Childs) and I hope that support will enhance for this feature.**

- **We have setup a site to coordinate development of Structural Biology applications on the Grid**
  - http://sci.cnb.uam.es/sbg/
- **Open to everybody**
- **You are welcome to join the team**

Acknowledgments

- **I want to give thanks to:**
  - CNB (Centro Nacional de Biotecnología).
  - CERN/ARDA group.
  - Site Administrators, Helpdesk support, Middelware and another people involved.

    (For their support and patience)