e-Science Applications in the EELA Grid

Rafael Mayo
CIEMAT
EGEE 07 Conference
Budapest, 1-5.10.2007

www.eu-eela.org

Information Society and Media

OUTLINE

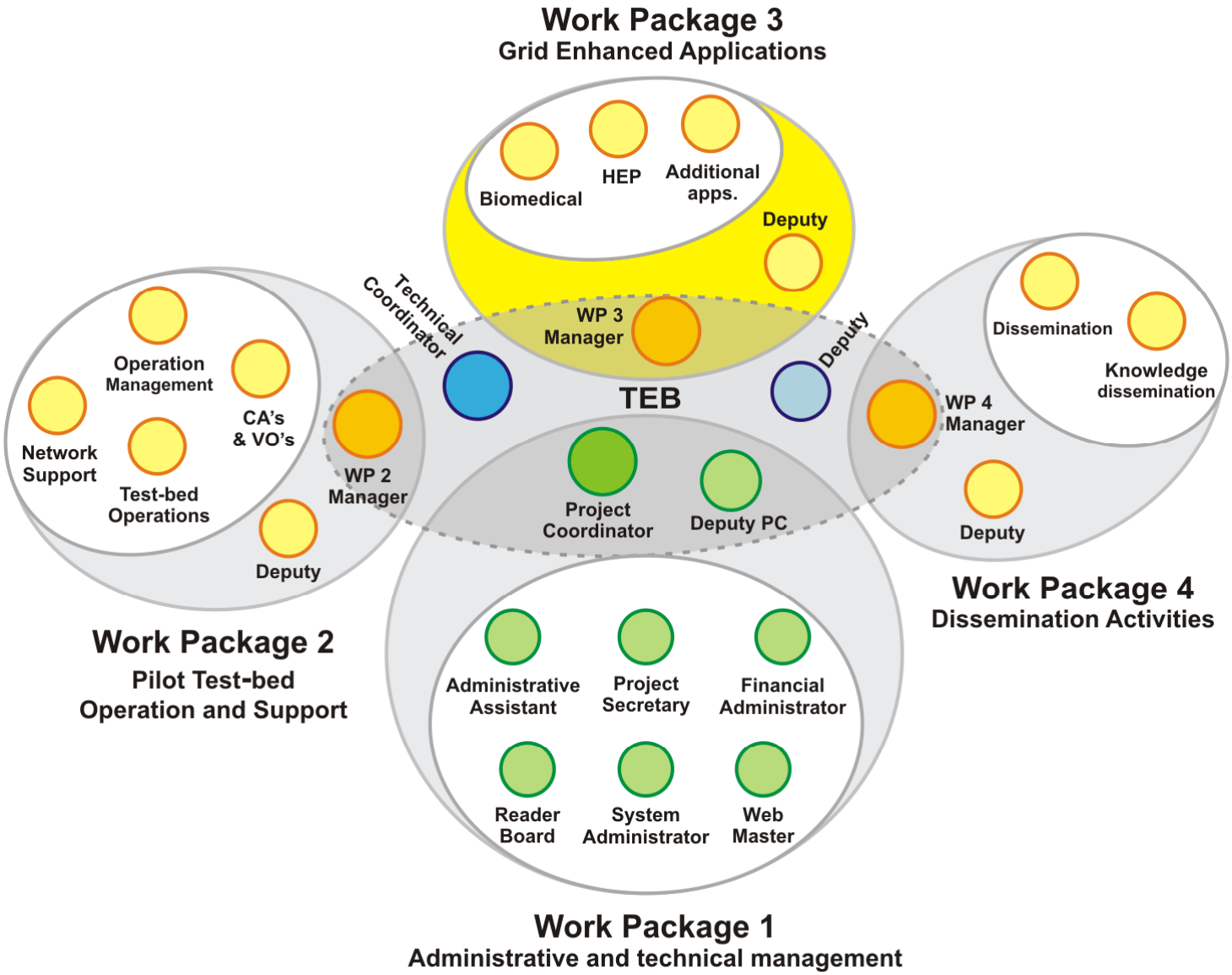
- **Applications Work Package**

- **EELA Applications**
  - Biomed
  - HEP
  - E-Learning
  - Climate

WORK PACKAGE 3

- **WP3: Identification and support of Grid enhanced applications**
  - Coordinated by CIEMAT
  - Identifies, selects and customizes relevant applications and tools suitable for the Grid dissemination process in:
    - T3.1 Biomedicine (CIEMAT, CUBAENERGIA, ULA, UNAM, UPV)
    - T3.2 High Energy Physics (CERN, CIEMAT, UFRJ, UNAM, UNLP, UTFSM)
    - T3.3 Additional Applications
      - *e-Education (CECIERJ/CEDERJ, CIEMAT, CUBAENERGIA, UFRJ, UNAM)*
      - *Climate (UC, UFCG, UDEC, SENAMHI)*
  - Aims at being the place of information exchange between already gridified applications and future ones.

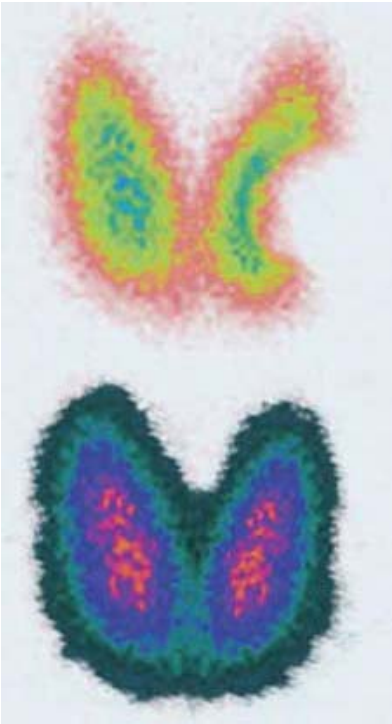
STRUCTURE


Work Package 3
Grid Enhanced Applications
Biomedical
HEP
Additional apps.
Deputy
WP 3 Manager
Technical Coordinator
Deputy
TEB
Operation Management
CA's & VO's
Network Support
Test-bed Operations
Deputy
WP 2 Manager
Dissemination
Knowledge dissemination
WP 4 Manager
Deputy
Project Coordinator
Deputy PC
Work Package 2
Pilot Test-bed Operation and Support
Work Package 4
Dissemination Activities
Administrative Assistant
Project Secretary
Financial Administrator
Reader Board
System Administrator
Web Master
Work Package 1
Administrative and technical management

Biomed

GATE

- **The interest of the LA community is leaded by CUBAENERGÍA**
  - It is focused towards two main oncological problems*:*
    - Thyroid Cancer.
    - Treatment of Metastasis with P^{32}.
  - 9 centers in Cuba are interested (*5 Hospitals* and *4 Oncological Centers and Institutions*)



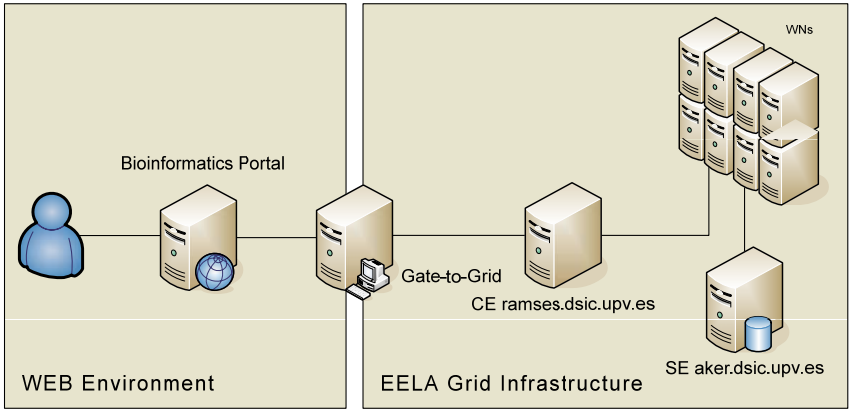
- **Installed in 4 EELA sites**
  - 62 processors
- **Installed in the Cuban Stand alone Grid**
  - 4 sites
  - GILDA framework

WISDOM

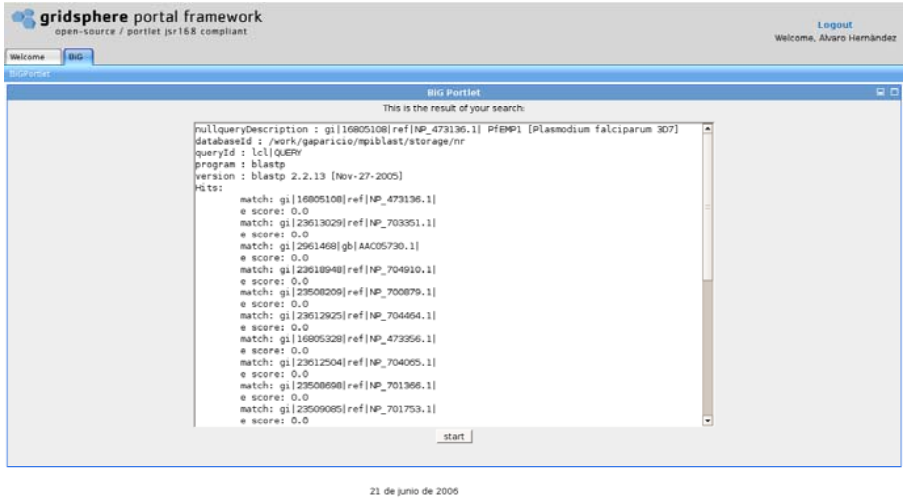
- **WISDOM can be executed on the EELA grid infrastructure.**
  - UPV already participated in previous Data Challenge on Malaria and Avian Flu.
- ***An Experiment was prepared jointly by UPV and ULA*.**
  - Two targets were accepted in *Plasmodium vivax Malaria* in the WISDOM Data Challenge-II
  - One has been completely docked (100%) in the EELA infrastructure
    - 2422 jobs
    - 53 GBytes
    - 228 CPU·days
    - Test the EELA infrastructure
- ***Next Steps will be*:**
  - Dock the second target
  - Analyze the first target
  - Inclusion of new targets in new data challenges.

BiG

- A Grid-enabled implemented to interface the mpiBLAST software
  - Bulk submission of simultaneous searches on several sequences
  - Users access the service through a Web Portal (CeCalcULA / UPV)
  - 2 Apps can be used (BLAST2GO / Basic stand-alone for testing)
  - Support to searching simultaneously on multiple databases
  - Access to resources for authenticated and authorized users (MyProxy server)
  - Robust: The system is Fault Tolerant on both the server and the client
- Access to the EELA Grid is performed through the *Gate-to-Grid* (an EELA Grid Node which provides a WSRF-Based Web Interface)



- In 1 year there were studies on
  - complete genome of the *Plasmodium falciparum* for the identification of DHFR antigenic proteins
  - Citric fruits gene databases
- 836 jobs executed.
- 263.4 CPU·days consumed.
- 4.6 CPU hours per job (1.1 CPU·hours of standard deviation)



- Current work in the Application
  - Sequential BLAST

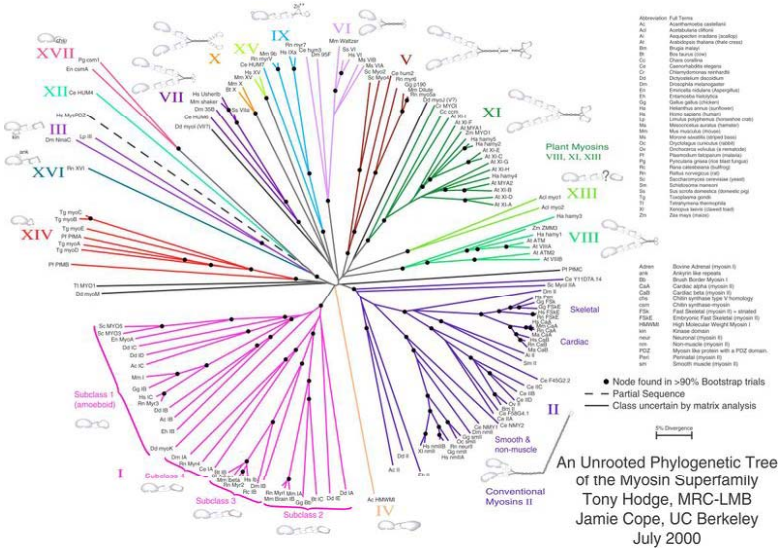
- Current experiment recently started
  - A large experiment on the alignment of different fragments of the genome of 20 different species of bacteria
  - It is not possible to make them grow isolated
    - EGEE (800) and EELA (100) processors
    - This experiment will require about 3 CPU years
    - Around 10K jobs of less than 4 hours each
    - Storage needs of 500 Mbytes of local storage per computing resource (temporally)
    - Space required on the SEs will be in the order of few GBytes in total
  - The objective is to reach 30K sequences per day → 1 month

PHYLOGENY

- **A Phylogeny is a Reconstruction of the Evolutionary History of a Group of Organisms.**

- **A Grid Service is being Developed to Run a Parallel Version of MrBayes from the Bioinformatics Portal (similar to BiG)**
  - Several sessions can be managed by the portal simultaneously
  - The results can be exported in different formats
  - Management of the VOMS credentials in the MyProxy repositories

- **But…**
  - Scalability is limited to a factor of 8 for each run
  - Executions in MrBayes are longer than in BLAST



PHYLOGENY

- **Successful previous tests with a paralell MrBayes vesion in a Grid-service**
  - Directed Acyclic Graph (DAG) technique
  - 121 sequences from the Papillomavirus were used as input for comparison
  - 13 sequences from gene fragment in Human Immunodeficiency Virus type 1 (HIV-1)
- **Some statistics with the Grid service developed by EELA**
  - The service was made available on June 07
  - Up to now, the CPU consume has been of 965 CPU hours in 24 jobs
  - It is important to outline that 97% of this time has been invested in two main jobs!
- **Future plans**
  - CLUSTAL-W are more scalable, although less accurate…
  - But for the classification of species at larger scope can be of interest

EMBOSS

- European Molecular Biology Open Software Suite

- Ported to the Grid by UNAM
- The databases from the tool will be stored in the Storage Elements and the LFC by means of the GFAL library
- The huge databases from NCBI are still to be copied
- Once this problem will be overcome, the RPMs will be distributed

- For a near future, a web service will be implemented

GAMOS / MIRaS

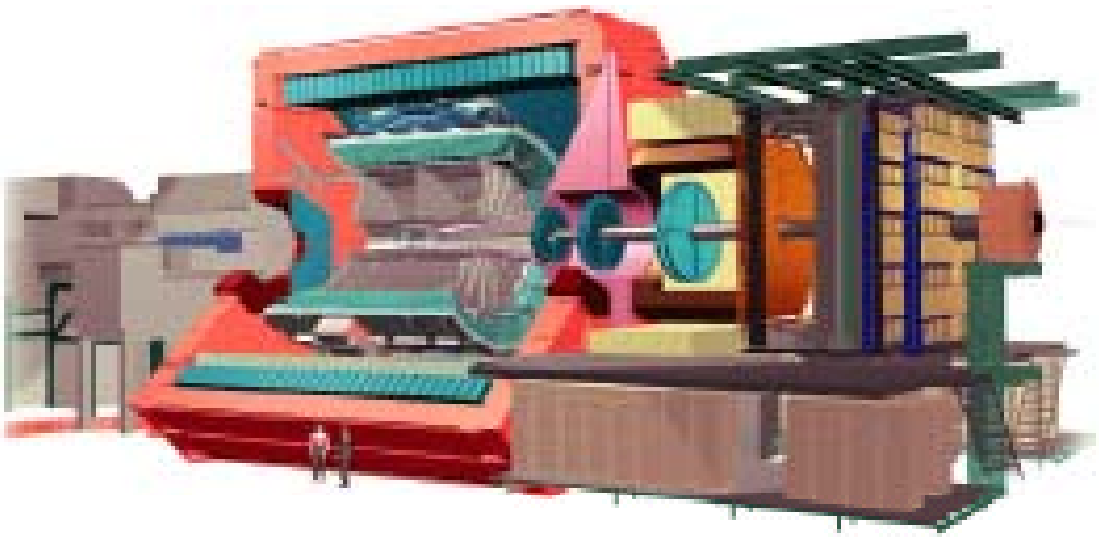
- **New tools developed for the simulation of the effects of the radiation in the human body**
- **Planned to offer to the community open tools to design medical imaging systems, verification and planification of treatment diseases**
- **They can adapt themselves to the available number or CPUs**
- **Contacts with**
  - Univ. Cayetano Heredia (Peru)
  - CINVESTAV (Mexico)
  - Univ. Tarapacá (Chile)
  - Puerta de Hierro Hospital (Spain)

HEP

– Initial applications

- **ALICE**

  To study the physics of strongly interacting matter at extreme energy densities, where the formation of a new phase of matter, the quark-gluon plasma, is expected.



ALICE IN EELA

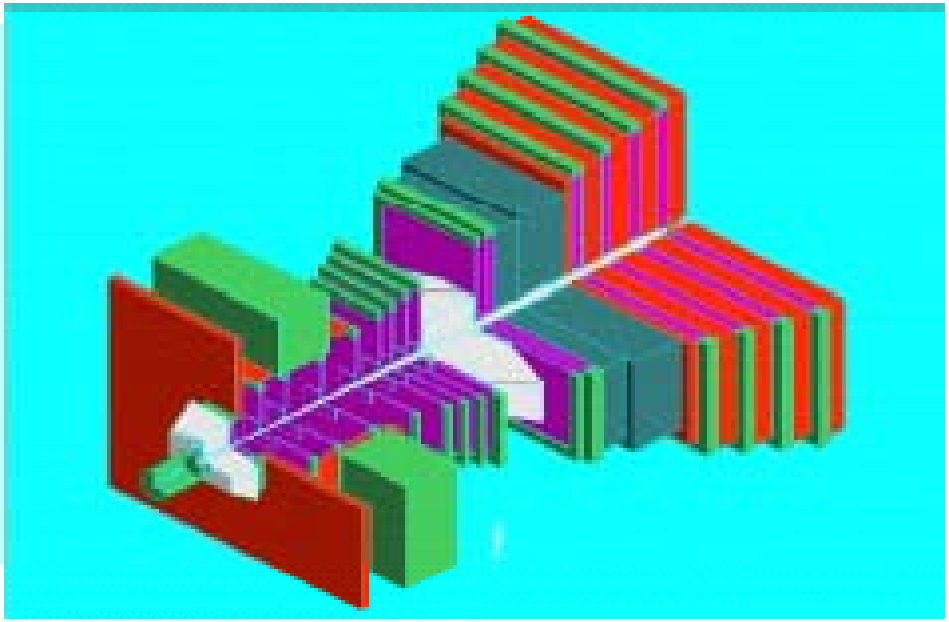
- **Three groups**
  - INFN-Catania
  - CIEMAT
  - UNAM
- **Four resource centres set up for ALICE jobs**
  - INFN-Catania
  - CIEMAT Madrid
  - CIEMAT Trujillo
  - UNAM

- **A total of 59000 jobs done on EELA sites in 2006**
- **Around 55000 successful jobs have been produced in the first semester of 2007**

– Initial applications

- **LHCb**

  To full investigate the CP violation in the Bd and Bs systems, to possibly renew the new physics beyond the standard model.



LHCb IN EELA

- **One group**
  - UFRJ
- **Three resource centers support LHCb**
  - INFN-Catania
  - CIEMAT
  - UFRJ
- **During 2006 EELA sites run 0.26% of the LHCb production**
  - It means around 44000 CPU hours and more than 4000 jobs
- **In the first semester of 2007 the percentage has raised to 0.29%**
  - 44558 CPU hours or almost 5500 jobs, i.e., 1.2% of the whole LHCb jobs production
- **The LHCb computing strategic management has decided to only support sites registered into the EGEE GOC Database**

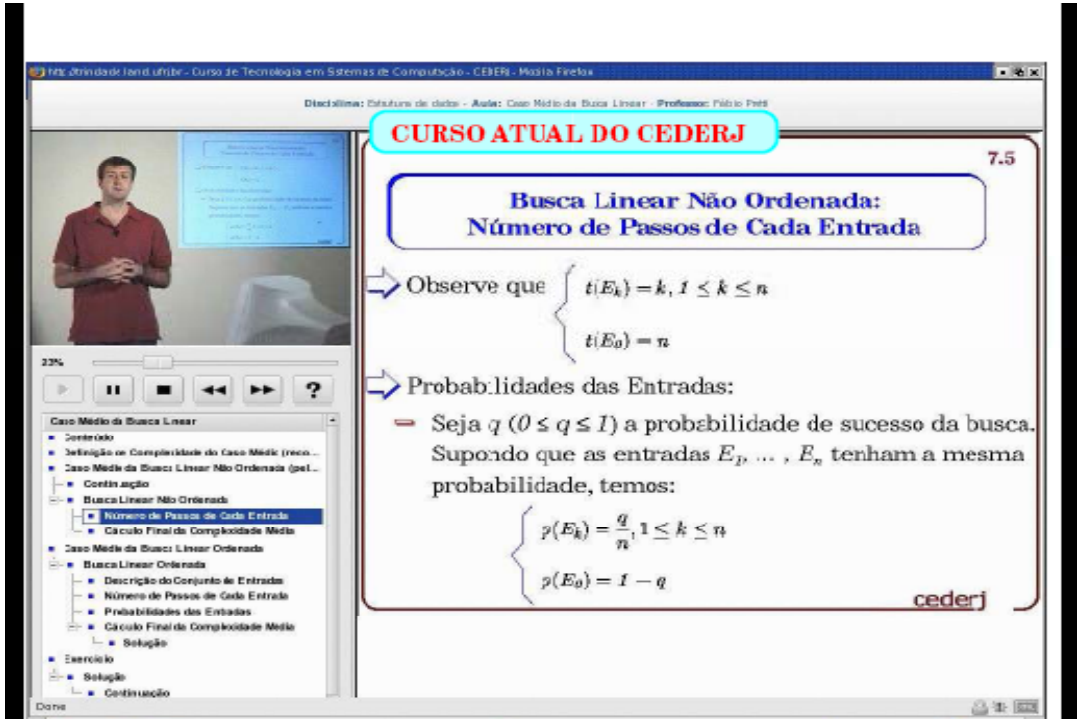
NEW HEP APPLICATIONS

- **Applications of interest to EELA partners and other communities in Latin America that have been improved their Grid feedback by means of EELA**
  - Other LHC application
    - ATLAS
      - *UNLP and UTFSM are setting up their infrastructure*
      - *It will be included in EELA-2*
  - New projects
    - Pierre Auger Observatory
      - *INFN-Catania, LIP, UFRJ, UNAM, and UNLP*
      - *An Auger VO and some key applications have been set up*
      - *It will be included in EELA-2*

e-Learning

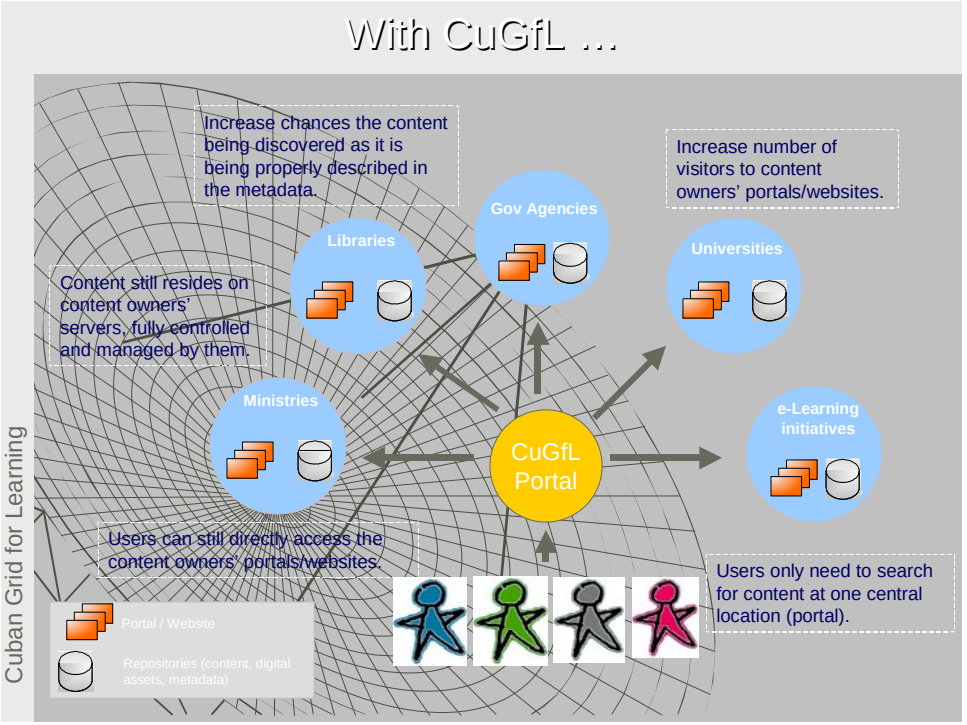
VoD

- Ported to the Grid by means of VO-Boxes
- Storage servers are running on the storage elements
- Multimedia server continues to manage requests
- Video streams continues to be sent to clients (UDP)
- Backup of classes stored on the grid
- Necessity of running with SL4 → Distribution to partners



CuGfl

- **Based on Moodle**
- **Course management system to produce web-based courses that support a social constructionist framework of education**
- **2 main modules**
  - Job Management
  - Authentication
- **Web Portal**
- **2 courses**
  - EELA Grid
  - Renewable Energies



LEMDist

- **Planed Services**
  - Access to distributed computer enhanced instrumentation
  - Remote access to simulation and modelation capabilities with high performance computing support
  - Interactive visualization
  - Distributed data analysis with access to data base systems
  - Experiment repository system.

Food Engineering

NMR

Electrophoresis

cns_solve

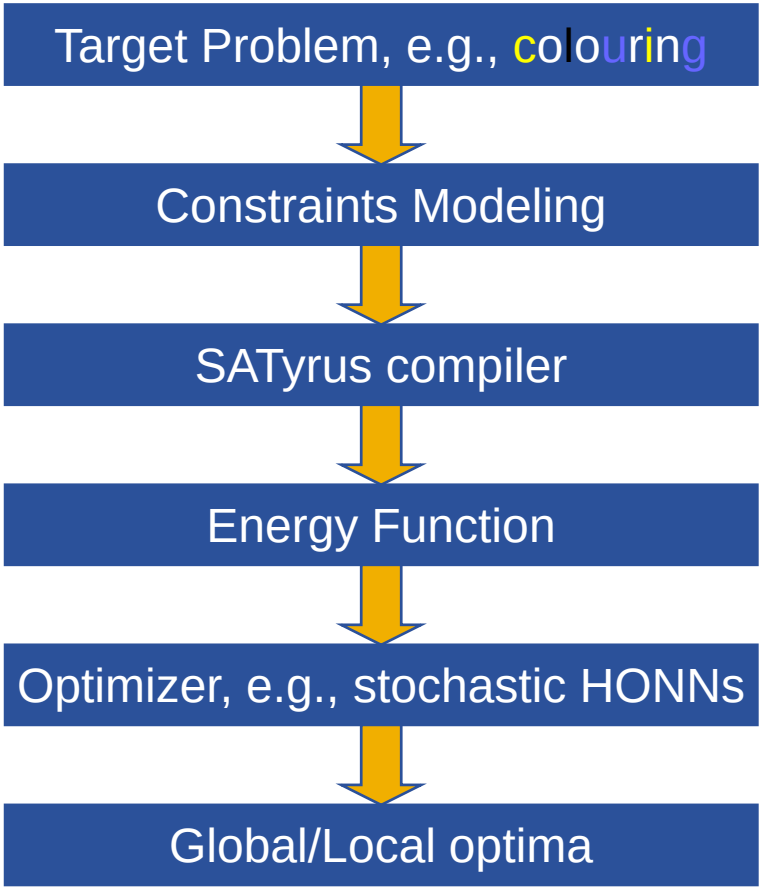
SATyrus

SATyrus:

- SATisfiability-based, neuro-symbolic architecture;

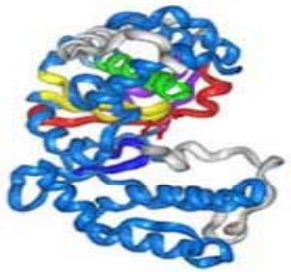
G-SATyrus

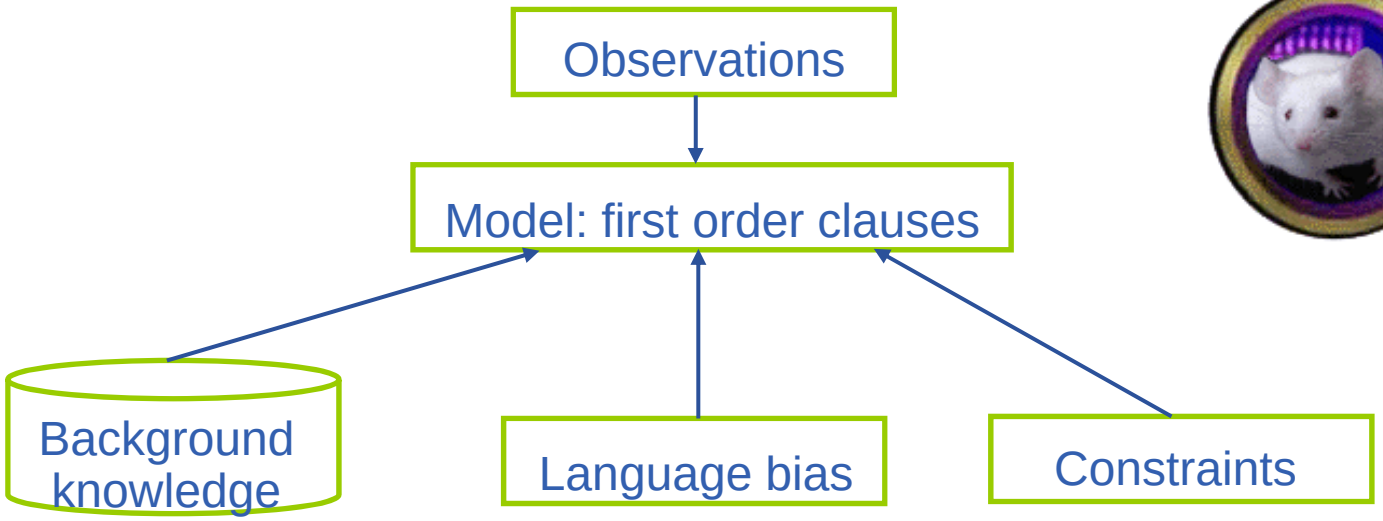
- Tested successfully
- Multi-start random searches launched through the CEs
- Monitoring by means of the gLite interface
- A new solver must be implemented in order to be more efficient in real problems



PILP

- **Use of Inductive Logic Programming to extract relevant knowledge from structured data**
- **Used to predict carcinogenesis in rodents**
  - A whole experiment still to be done





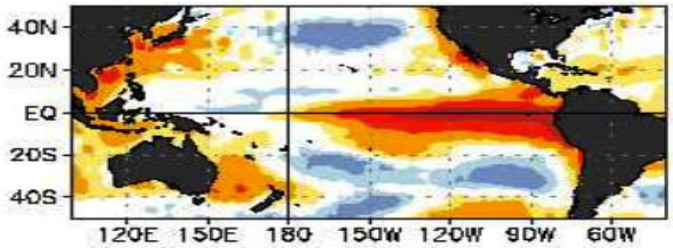
Climate

CLIMATE CHALLENGE

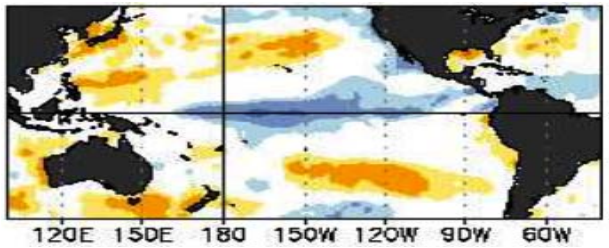
- **Goal: Predict Local Impacts of "El Niño" in Latin America**

  A challenging problem for the climate community, with **huge** socio-economical impact in Latin America.

Anomalous heating



Anomalous cooling



- **GRID helps to** share computing resources, heterogeneous data, as well as know-how in a user-friendly form.

- **A new integrated climate application** developed in EELA from the scratch, with no similar counterpart in any other Earth Science/Climate EU Project.

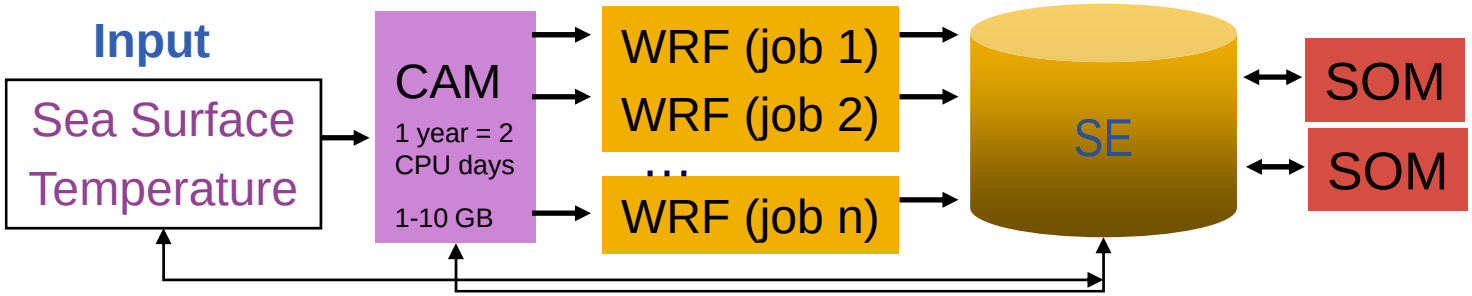
CLIMATE APPLICATIONS

Three applications have been identified (climate sequence):

- Global atmospheric circulation model (**CAM**) …….. ***Deployed !***
- Regional weather model (**WRF**) …….. ***Deployed !***
- Datamining clustering tools (SOM) …….. *Work in progress*

This sequence poses several computational challenges

Nontrivial dependent relationships among the applications.

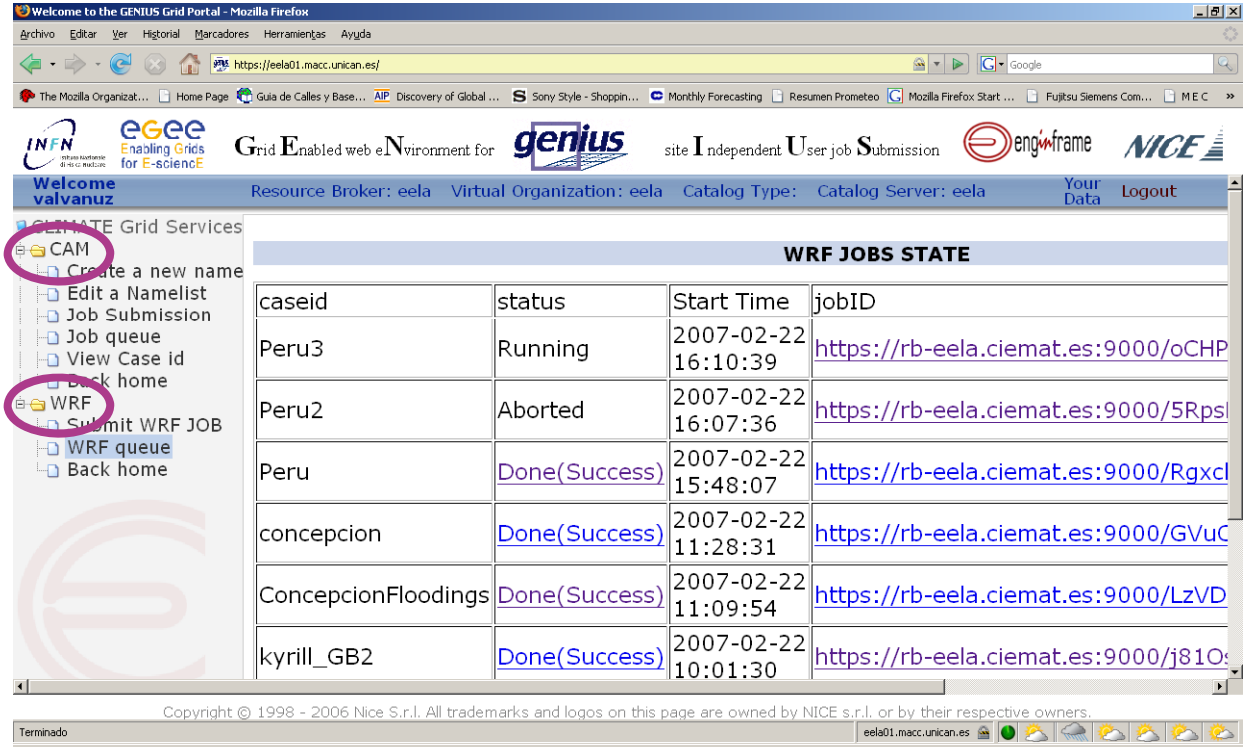


This sequence of jobs demands middleware solutions for:

- Preparing and submitting dependent jobs / data sharing (workflow).
- Restarting interrupted experiments.
- Manage metadata (for datasets and application status).

CURRENT STATUS

- Some work still remain to couple the output of CAM as boundary condition for WRF
- Using GENIUS to interact with the applications (CAM+WRF)



WORK IN PROGRESS

- Study the Sensitivity of El Niño Precipitation to Sea Surface Temperature (SST) Perturbations
- Deploy an appropriate workflow management system to run this scientific challenges → implementation of a proper solution for the workflow management using metadata catalogs (AMGA)
- Creation of a database in AMGA to store and manage the information required for the experiment
  - This AMGA structure defines "status flags" to monitor each of the jobs and produces the required next step in order to get every job successfully finished in an unsupervised form

SEGHIDRO

- Deployed by the Universidade Federal de Campina Grande
- Aims at improving the water management of the Brazilian Northeast
- Enables, through a grid portal, collaborative work via the coupling of computer models (BRAMS) providing access to massive grid-based computer resources
- The portal uses an underlying grid infrastructure named “The OurGrid Community”

- Interoperability between EELA and OurGrid middlewares!

Other Applications

Volcano Sonifications

VOLCANO SONIFICATIONS

- **Currently no definitive method to predict the eruption of a volcano has been discovered or implemented (yet)**

- **Some of the calculations have been performed in the EELA e-Infrastructure.**

Useful Information

WP3 web page:

http://www.eu-eela.org/eela_wp3.php

WP3 documents:

http://documents.eu-eela.org

Contacts

WP3 contacts

Vicente Hernández (Biomed)

vhernand@dsic.upv.es

Lukas Nellen (HEP)

lukas@nucleares.unam.mx

Inês Dutra (e-Learning)

ines@dcc.fc.up.pt

José Manuel Gutiérrez (Climate)

manuel.gutierrez@unican.es

Rafael Mayo

rafael.mayo@ciemat.es

Thanks for your attention!