Applications on the EGEE Grid infrastructure

Slides from EGEE project members

www.eu-egee.org

Information Society and Media

Overview

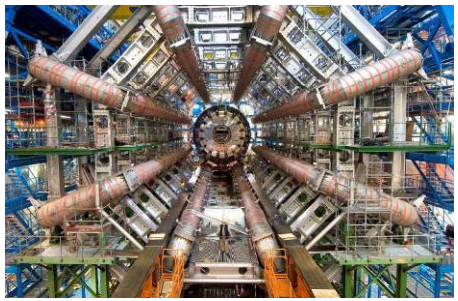
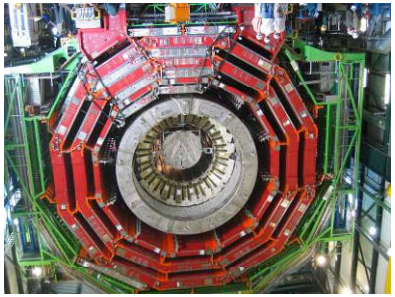
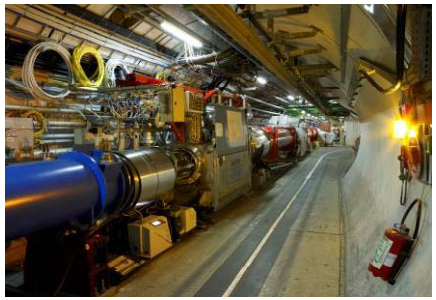
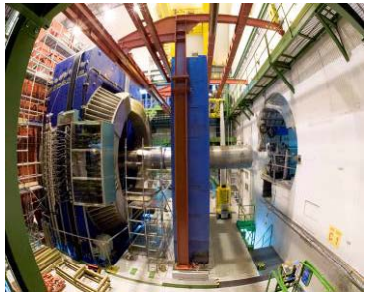
- **Two “pilot” application areas in EGEE**
  - High Energy Physics
  - Biomedical
- **Examples of the growing number of diverse communities**
- **What’s being done?! A classification of applications**
- **Challenges for application developers**

- **NOTE**
  - Not giving details on each application but look for possible patterns and similarities with your own interests – perhaps in different disciplines
  - Grid middleware supports commonly met requirements so techniques will also cross between different communities

High Energy Physics

- **High Energy Physics is a pilot application domain for EGEE**
  - Large datasets
  - Large computing requirements
  - → Major need for Grid technology to support distributed communities

- **Support for Large Hadron Collider experiments through LHC Computing Grid (LCG)**
  - ATLAS, CMS, LHCb, ALICE

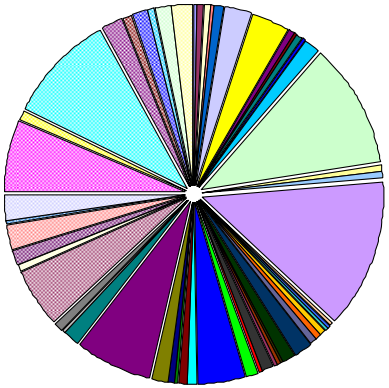





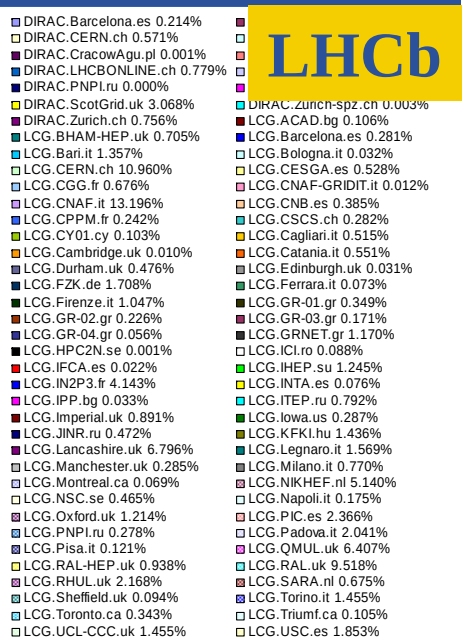
- **Also support for other major international HEP experiments**

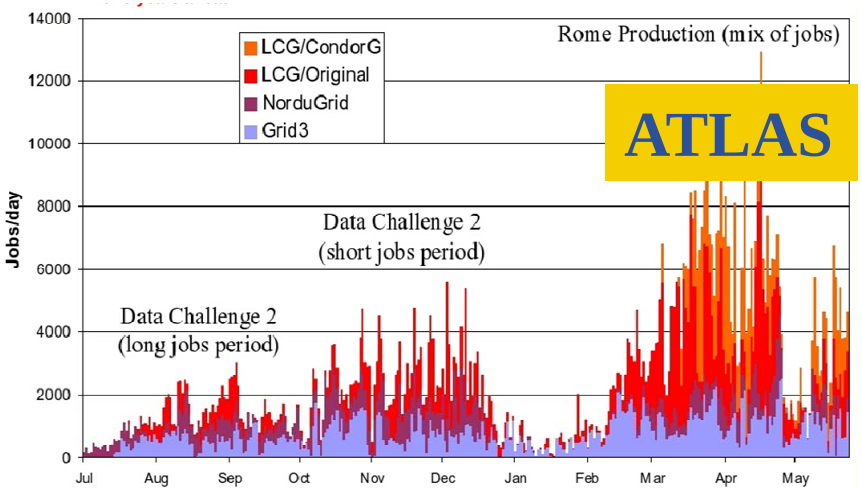
HEP success stories

- **Fundamental activity in preparation of LHC start up**
  - Physics
  - Computing systems
- **Examples:**
  - LHCb: **~700 CPU/years in 2005** on the EGEE infrastructure
  - ATLAS: **over 20,000 jobs per day**
    - Comprehensive analysis: see S.Campana et al., "Analysis of the ATLAS Rome Production experience on the EGEE Computing Grid", e-Science 2005, Melbourne, Australia
  - A lot of activity in all involved applications (including as usual a lot of activity within non-LHC experiments like BaBar, CDF and D0)

CPU used: 6,389,638 h
Data Output: 77 TB

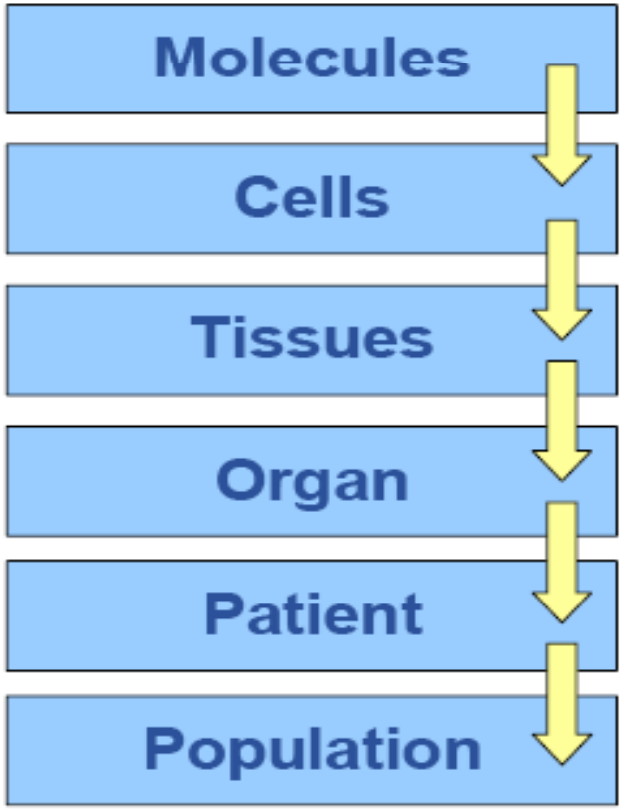






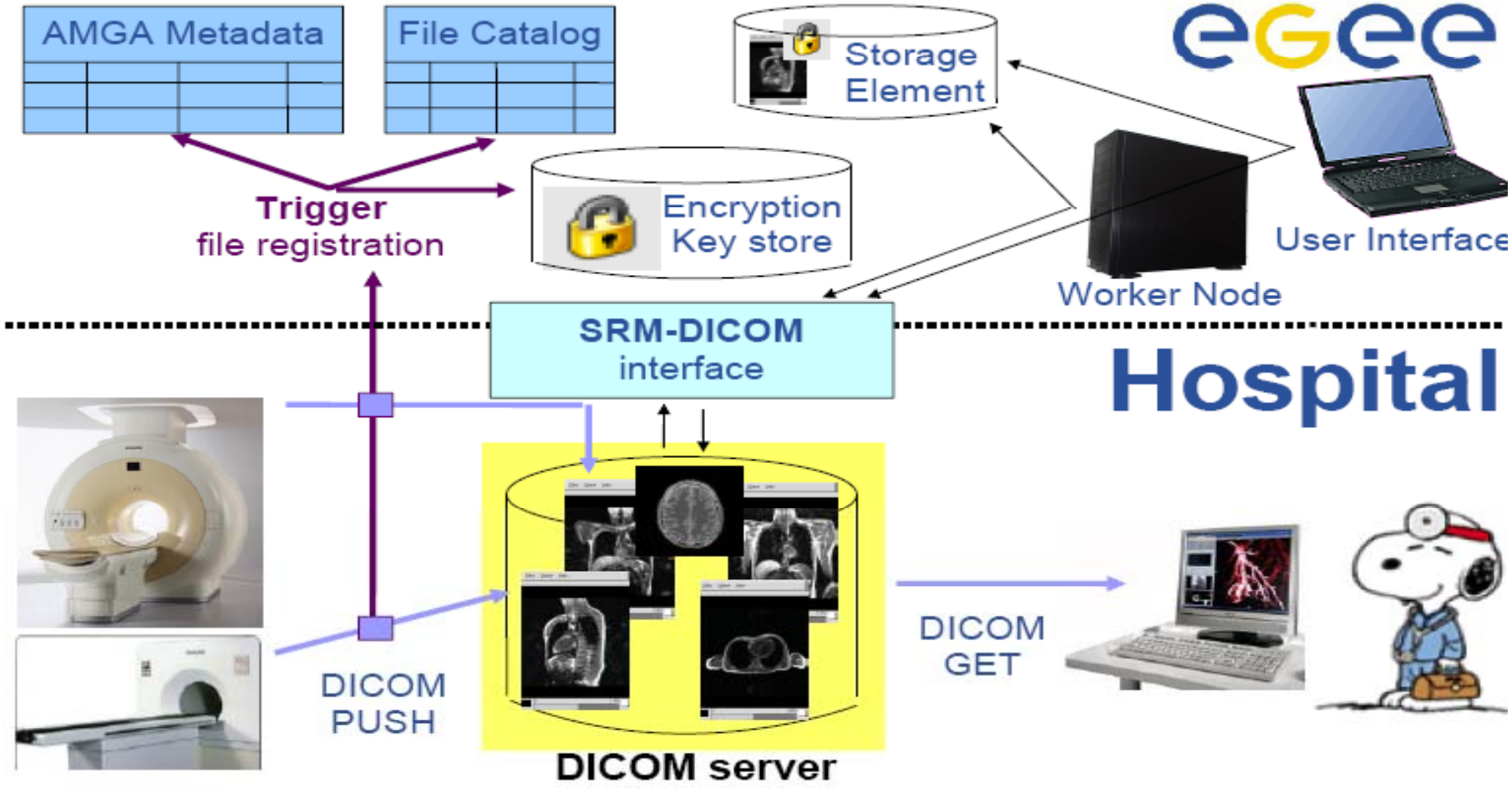
Biomedical applications

- **Bioinformatics**
  - Genomics
  - Proteomics
  - Phylogeny…
- **Medical imaging**
  - Medical imaging
  - Computer Aided Diagnosis
  - Therapy planning
  - Simulation…
- **Life sciences**
  - Drug discovery
  - Epidemiology
  - …



Biomedical community and the Grid, EGEE User Forum, March 1st 2006, I. Magnin

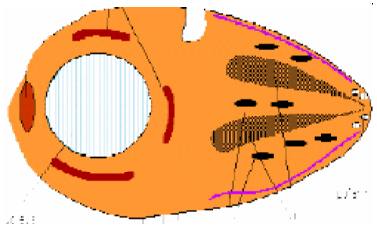
Data management – medical images



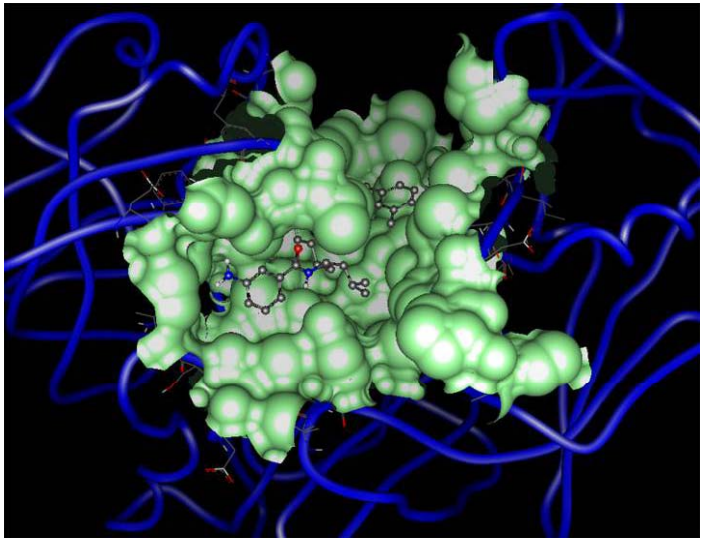
Biomedical community and the Grid, EGEE User Forum, March 1st 2006, I. Magnin

Applications Example: WISDOM

- **Grid-enabled drug discovery process for neglected diseases**
  - *In silico* docking
    - compute probability that potential drugs dock with target protein
  - To speed up and reduce cost to develop new drugs

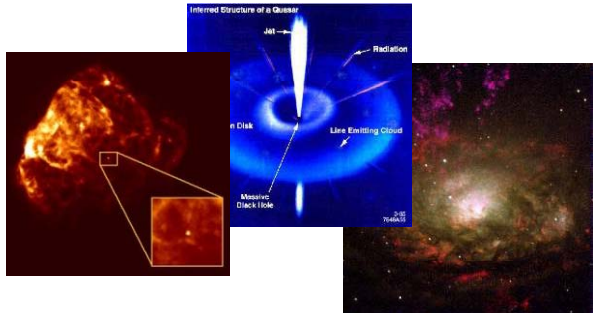


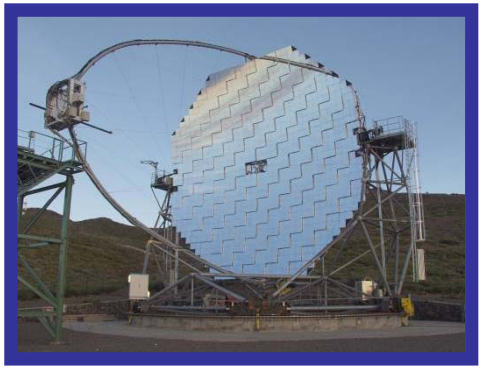
- **WISDOM (World-wide In Silico Docking On Malaria)**
  - First biomedical data challenge
  - 46 million ligands docked in 6 weeks
  - 1TB of data produced
  - 1000 computers in 15 countries
    - Equivalent to 80 CPU years
- **Second data challenge on Avian flu in April 2006**
  - 300,000 possible drug components tested
  - 8 different targets
  - 2000 computers used for 4 weeks



Astroparticle physics: MAGIC

- **Major Atmospheric Gamma Imaging Cherenkov telescope (MAGIC)**
  - Origin of VHE γ-rays (30 GeV – TeV)
    - Active Galactic Nuclei (AGN)
    - Supernova Remnants
    - Unidentified EGRET sources
    - Gamma Ray Bursts
  - Huge hadronic background → MC simulations
    - to simulate the background of one night, 70 CPUs (P4 2GHz) need to run for 19200 days
  - Observation data are big too!



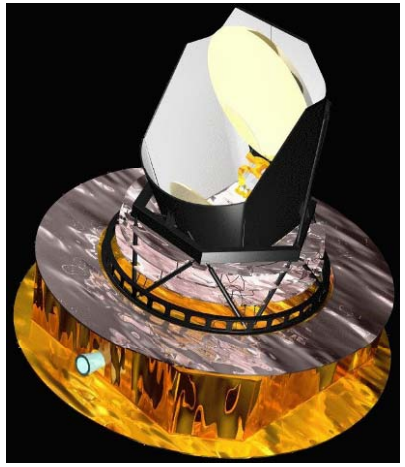


- **MAGIC Grid**
  - Use three national Grid centres as backbone
  - All are members of EGEE

- **Work to build a second telescope is currently in progress**

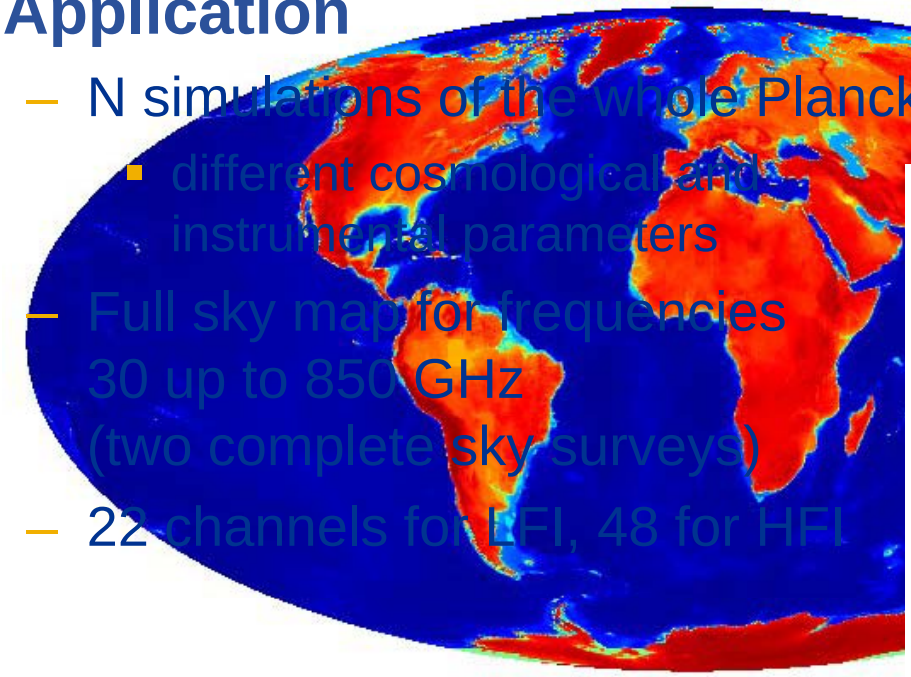
→ **Towards a virtual observatory for VHE γ -rays**

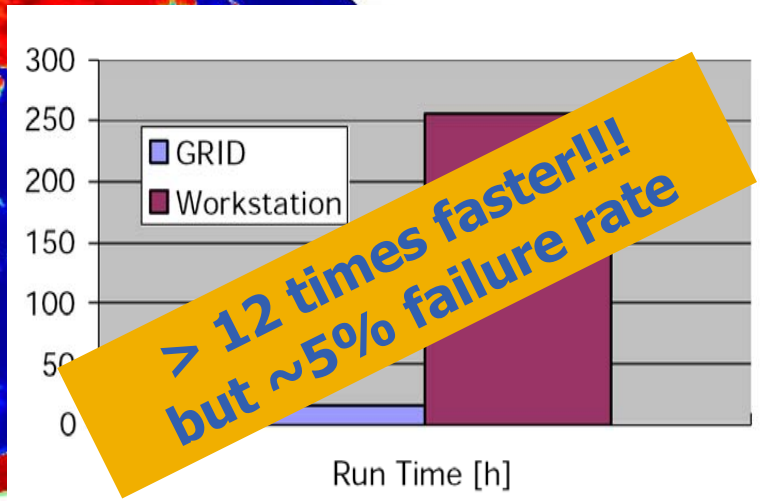
Astroparticle physics: PLANCK



- **PLANCK satellite mission**
  - Measure cosmic microwave background (CMB)
    - At even higher resolution than previous missions
  - Launch in 2008; duration >1 year

- **Application**
  - N simulations of the whole Planck/LFI mission
    - different cosmological and instrumental parameters
  - Full sky maps for frequencies 30 up to 850 GHz (two complete sky surveys)
  - 22 channels for LFI, 48 for HFI



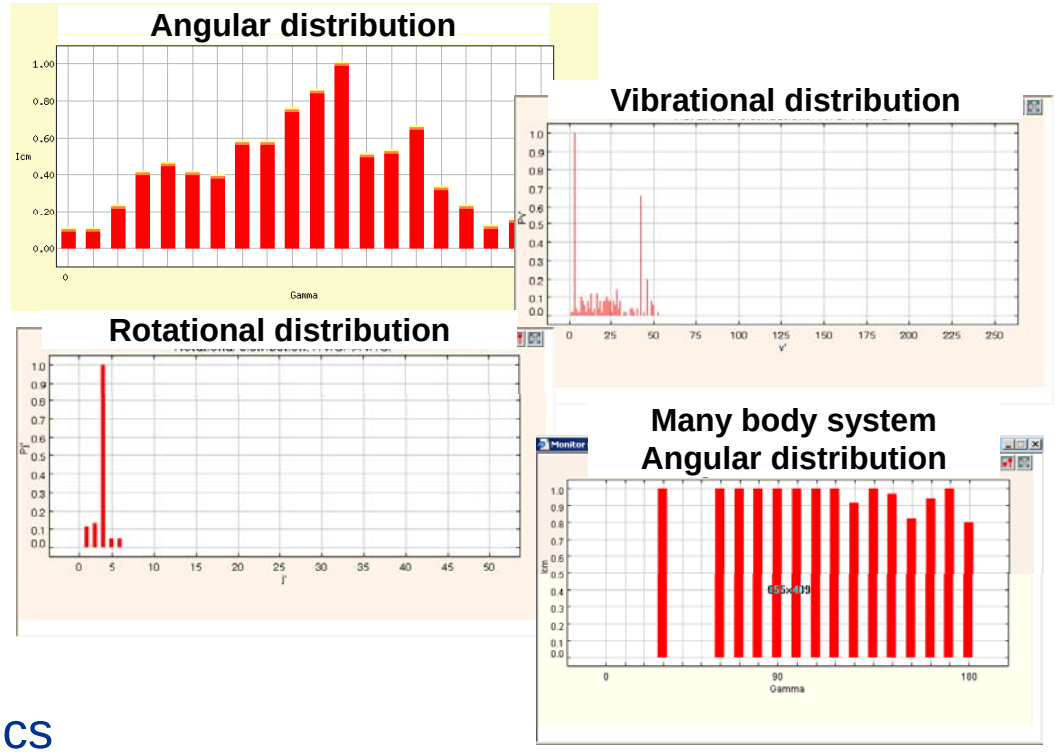


Computational Chemistry

- **GEMS (Grid Enabled Molecular Simulator) application**
  - Calculation and fitting of electronic energies of atomic and molecular aggregates (using high level *ab initio* methods)
  - The use of statistical kinetics and dynamics to study chemical processes

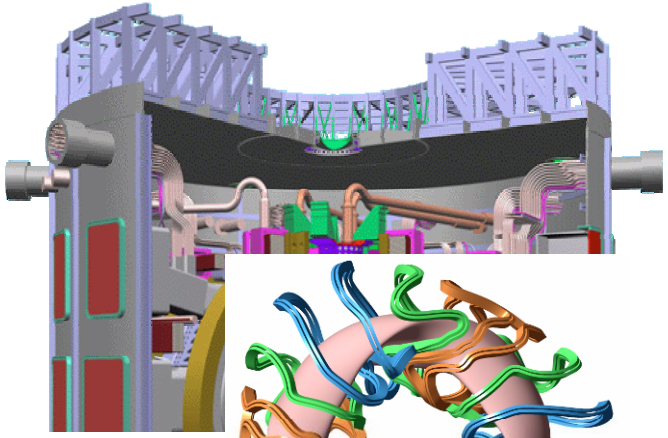
- **Virtual Monitors**
  - Angular distributions
  - Vibrational distributions
  - Rotational distributions
  - Many body systems

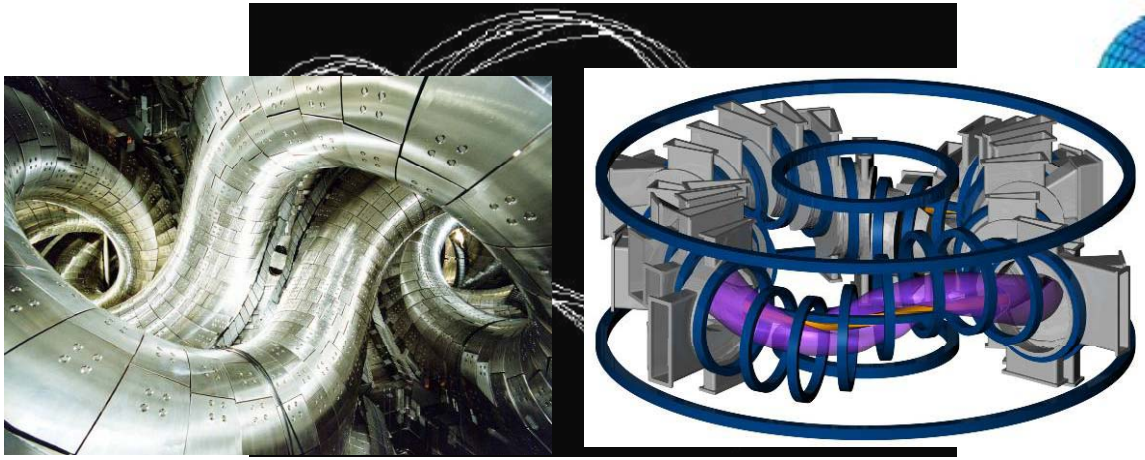
- **End-User applications**
  - Nanotubes
  - Life sciences
  - Statistical Thermodynamics
  - Molecular Virtual Reality



Fusion

- **Large Nuclear Fusion installations**
  - E.g. International Thermonuclear Experimental Reactor (ITER)
  - Distributed data storage and handling needed
  - Computing power needed for
    - Making decisions in real time
    - Solving kinetic transport
      → particle orbits
    - Stellarator optimization
      → magnetic field to contain the plasma

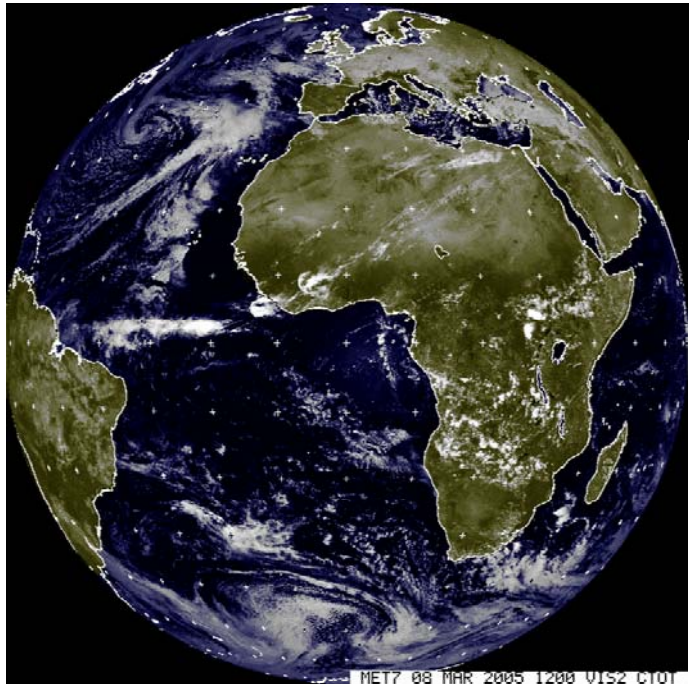




Earth Science Applications

- **Community**
  - Many small groups that aggregate for projects (and separate afterwards)

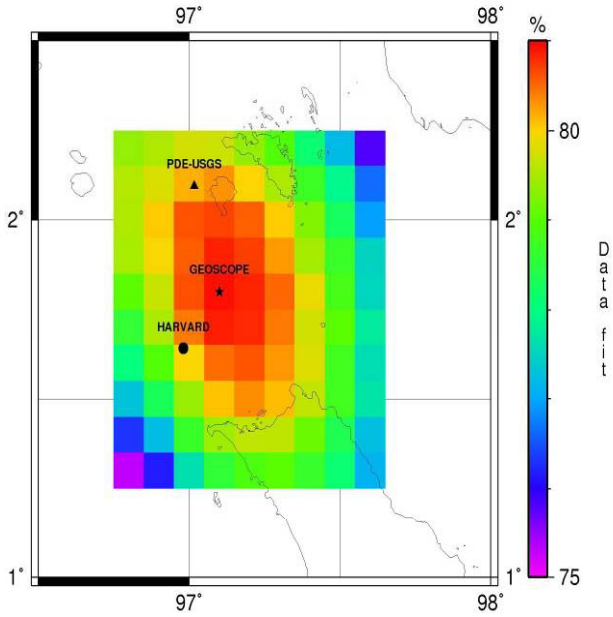
- **The Earth**
  - Complex system
  - Independent domains with interfaces
    - Solid Earth – Ocean – Atmosphere
  - Physics, chemistry and/or biology



- **Applications**
  - Earth observation by satellite
  - Seismology
  - Hydrology
  - Climate
  - Geosciences
  - Pollution
  - Meteorology, Space Weather
  - Mars Atmosphere
  - Database Collection

Earth Sciences: Earthquake analysis

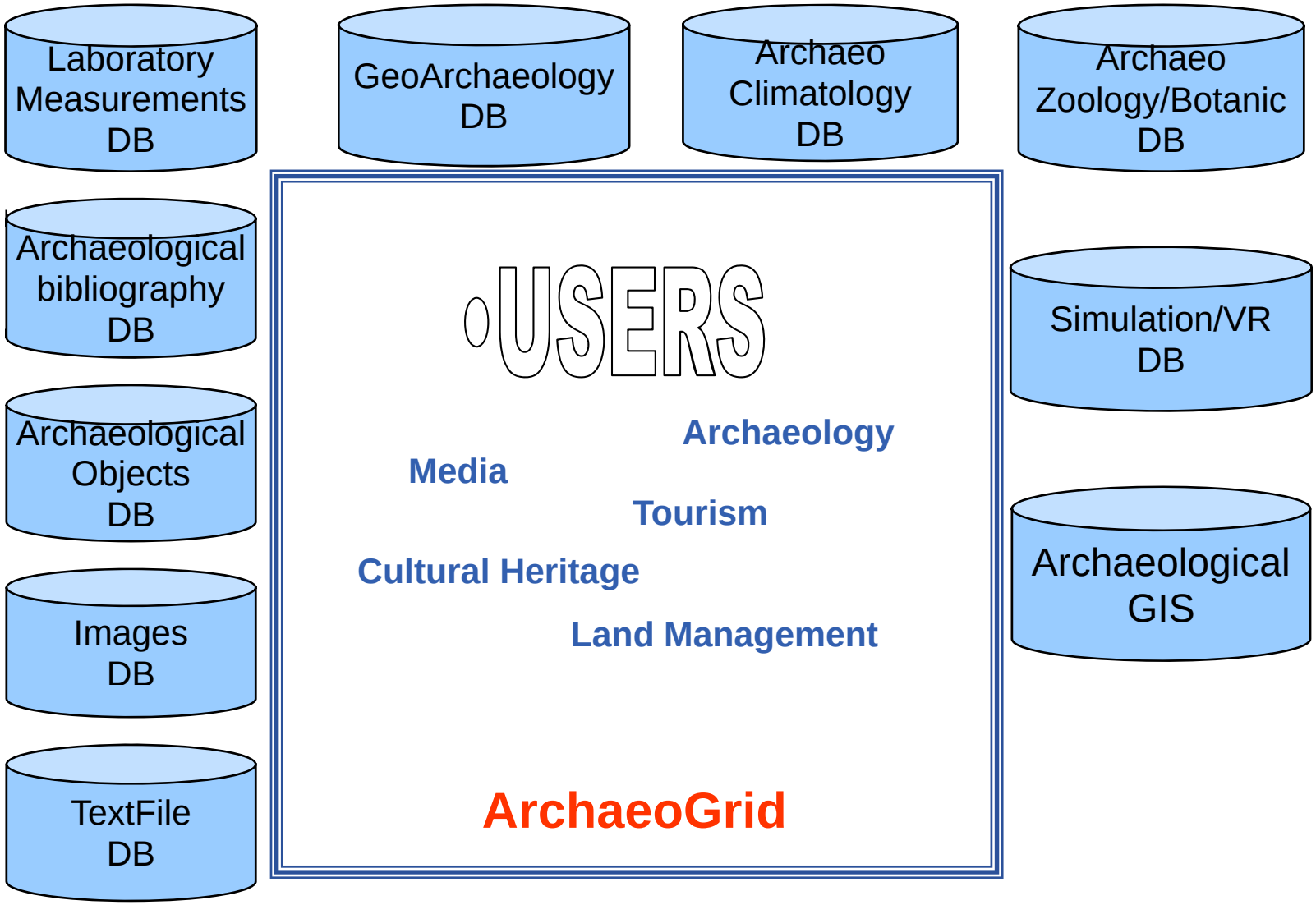
- **Seismic software application determines:**
  Epicentre, magnitude, mechanism
  → May make it possible to predict future earthquakes
  → Assess potential impact on specific regions



- **Analysis of Indonesian earthquake (28 March 2005)**
  - Data from French seismic sensor network GEOSCOPE transmitted to IPGP within 12 hours after the earthquake
  - Solution found within 30 hours after earthquake occurred
    - 10 times faster on the Grid than on local computers
  - Results
    - Not an aftershock of December 2004 earthquake
    - Different location (different part of fault line further south)
    - Different mechanism

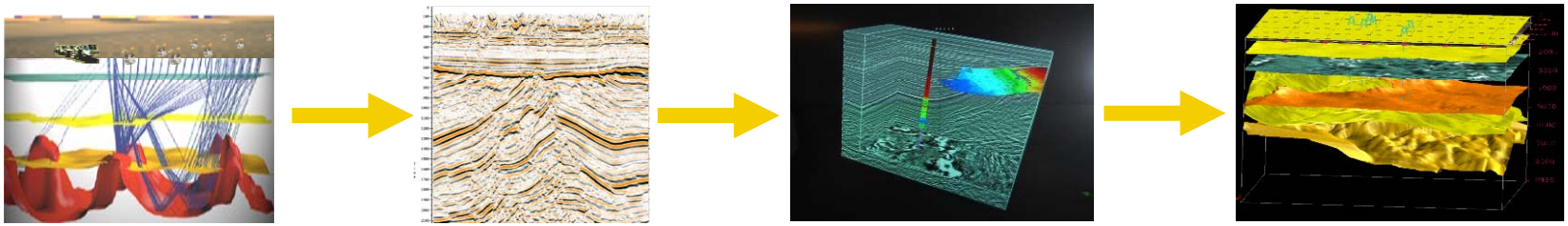
- **Rapid analysis of earthquakes is important for relief efforts**

ArchaeoGrid



Industrial applications

- **EGEODE**
  - Industrial application from Compagnie Générale de Géophysique running on EGEE infrastructure
    - Seismic processing platform
    - Based on industrial application Geocluster© used at CGG
    - Being ported to EGEE for Industry and Academia



- **OpenPlast project**
  - French R&D programme to develop and deploy Grid platform for plastic industry (SMEs)
  - Based on experience from EGEE (supported by CS)
  - Next: Interoperability with other Grids

EGEE User Forum

- **Annual event allowing users to share experiences and give feedback to EGEE**

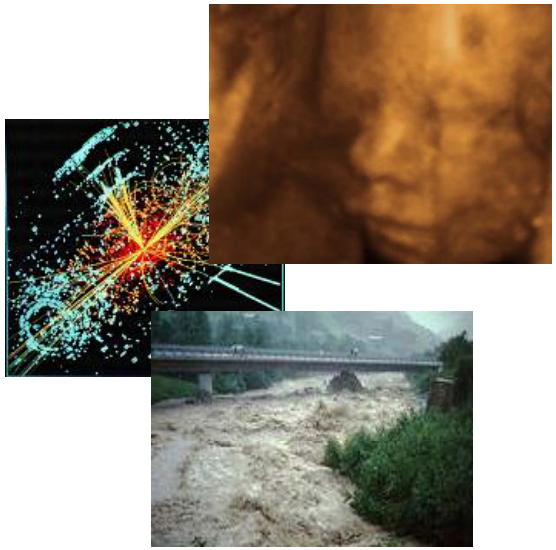
  Extremely successful and popular mechanism for:

  - Increasing interactions between users
  - Presenting what has been achieved using grid technology
  - Discussing problems and solutions

www.eu-egee.org/uf2
Enabling Grids for E-sciencE
EGEE USER FORUM
9-11 MAY 2007 MANCHESTER (UK)

Conclusion

- **>200 Virtual Organisations from 10 domains, many applications under evaluation**
- **Production-quality gLite middleware to support many different application groups**
- **User Forum: Annual event allowing users to share experiences and give feedback to EGEE**



Overview

- Two “pilot” application areas in EGEE
  - High Energy Physics
  - Biomedical
- Examples of the growing number of diverse communities
- **What’s being done?! A classification of applications**
- **Challenges for application developers**

- **NOTE**
  - Not giving details on each application but look for possible patterns and similarities with your own interests – perhaps in different disciplines
  - Grid middleware supports commonly met requirements so techniques will also cross between different communities

Characteristics of Grid Applications

C. Loomis (LAL-Orsay)

EGEE'06 Conference (Geneva)

25-29 September 2006

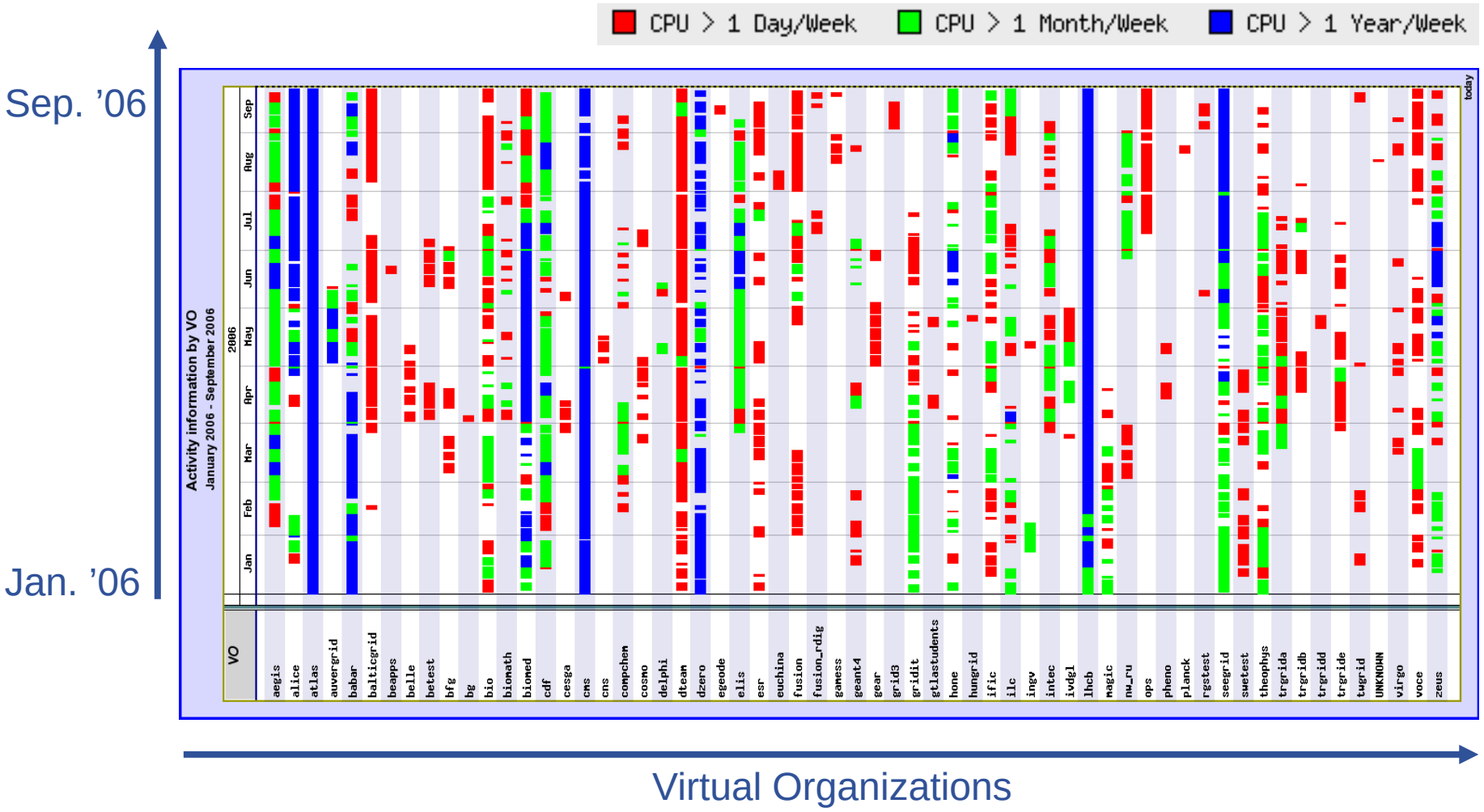
Outline

- **Status**
  - EGEE's users, applications, and virtual organizations
  - "Application Identification and Support" activity
  - Evolution: project, users, and needs
- **Grid Application "Families"**
- **Summary and Outlook**

Current Activity

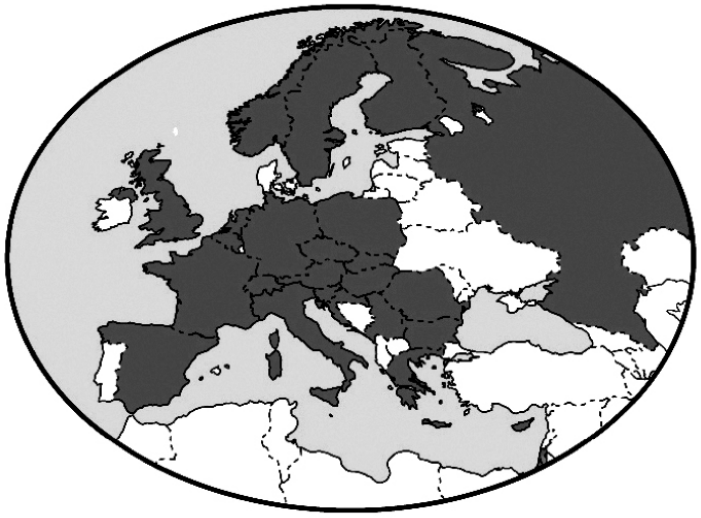
- **Routine and large-scale use of EGEE infrastructure to produce scientific results.**
- **VOs:**
  - 165+ VOs (90+ registered) using the grid
  - App. Deploy. Plan (https://edms.cern.ch/document/722131/2)
- **Domains:**
  - **High-Energy Physics**: LHC, Tevatron, HERA, …
  - **Biology**: Medical Images, Bioinformatics, Drug Discovery
  - **Earth Science**: Hydrology, Pollution, Climate, Geophysics, …
  - **Astrophysics**: Planck, MAGIC
  - **Fusion**
  - **Computational Chemistry**
  - **Related Projects**: Finance, Digital Libraries, …
  - **New areas**: nanotechnology, …

CPU Usage


CPU > 1 Day/Week
CPU > 1 Month/Week
CPU > 1 Year/Week
Sep. '06
Jan. '06
Activity information by VO
January 2006 - September 2006
2006
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
today
VO
aegis alice atlas auvergrid babar balticgrid beapps belle betest bfg bg bio biomath bioned cdf cesga cms cms compchem cosmo delphi dteam dzero egeode elis esr euchina fusion fusion_rdig gamess geant4 gear grid3 gridit gtlastudents hone hungrid ific ilc ingv intec ivdgl lhcb magic na_ru ops pheno planck rgstest seegrid swetest theophys trgrida trgridb trgridd trgride tugrid UNKNOWN virgo voce zeus
Virtual Organizations

NA4 Activity

- **Application Identification and Support (NA4)**
  - 25 countries, 40 partners, 280+ participants, 1000s of users
- **Support the large and diverse EGEE user community:**
  - **Promote dialog**: Users' Forums & EGEE Conferences
  - **Technical Aid**: Porting code, procedural issues
  - **Liaison**: Software and operational requirements
- **Need active participation:**
  - **Feedback**: Infrastructure, configuration, and middleware
  - **Resources**: Hardware and human



Evolution

- **Evolution of Project (2001–now):**
  - European DataGrid: R&D
  - EGEE: Re-engineering & Infrastructure
  - EGEE-II: Infrastructure & Re-engineering

larger grid ↓ more apps.

- **Evolution of Grid Users:**
  - **Focus**: Grid technology ⇒ Scientific results
  - **Goal**: Grid technology ⇒ Grid as a tool
  - **Experience**: IT experts ⇒ IT “minimalists”

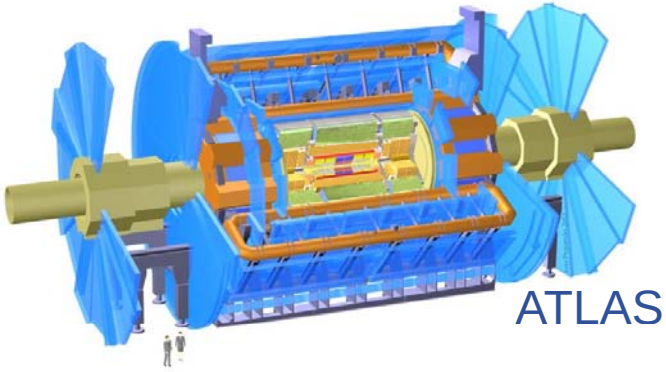
- **These changes are healthy, but…**
  - Rely less on IT competence of users.
  - More portable, more flexible middleware.

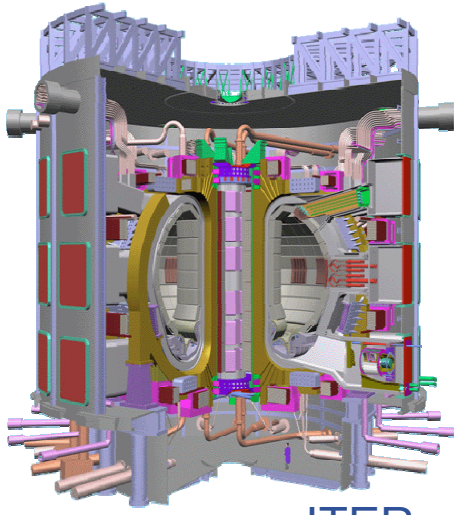
Application Families

- Simulation
- Bulk Processing
- Responsive Apps.
- Workflow
- Parallel Jobs
- Legacy Applications

Simulation

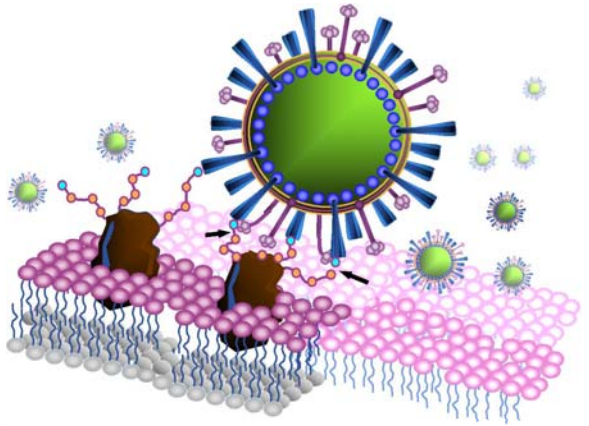
- **Examples**
  - LHC Monte Carlo simulation
  - Fusion
  - WISDOM—malaria/avian flu
- **Characteristics**
  - Jobs are CPU-intensive
  - Large number of independent jobs
  - Run by few (expert) users
  - Small input; large output
- **Needs**
  - Batch-system services
  - Minimal data management for storage of results


ATLAS


ITER

Drug Discovery

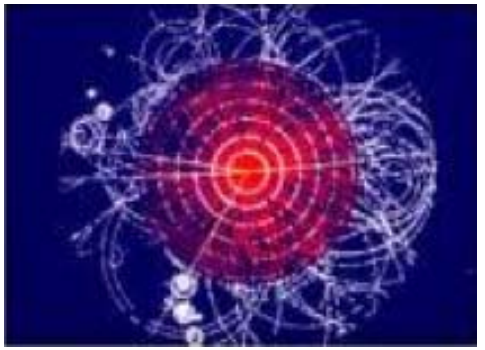
- **WISDOM focuses on in silico drug discovery for neglected and emerging diseases.**

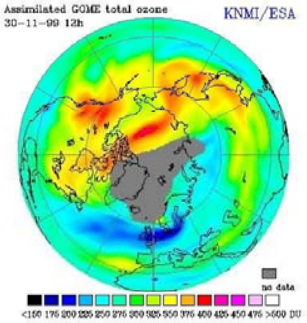


- **Malaria — Summer 2005**
  - 46 million ligands docked
  - 1 million selected
  - 1TB data produced; 80 CPU-years used in 6 weeks
- **Avian Flu — Spring 2006**
  - H5N1 neuraminidase
  - Impact of selected point mutations on eff. of existing drugs
  - Identification of new potential drugs acting on mutated N1
- **Fall 2006**
  - Extension to other neglected diseases

Bulk Processing

- **Examples**
  - HEP processing of raw data, analysis
  - Earth observation data processing
- **Characteristics**
  - Widely-distributed input data
  - Significant amount of input and output data
- **Needs**
  - Job management tools (workload management)
  - Meta-data services
  - More sophisticated data management



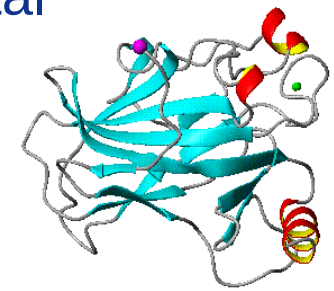


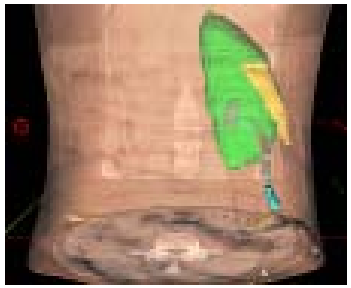
Responsive Apps. (I)

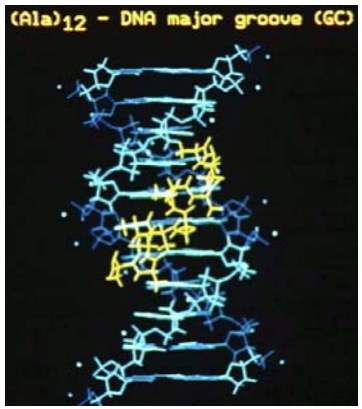
- **Examples**
  - Prototyping new applications
  - Monitoring grid operations
  - Direct interactivity
- **Characteristics**
  - Small amounts of input and output data
  - Not CPU-intensive
  - Short response time (few minutes)
- **Needs**
  - Configuration which allows “immediate” execution (QoS)
  - Services must treat jobs with minimum latency

Responsive Apps. (II)

- **Grid as a backend infrastructure:**
  - gPTM3D: interactive analysis of medical images
  - GPS@: bioinformatics via web portal
  - GATE: radiotherapy planning
  - DILIGENT: digital libraries
  - Volcano sonification



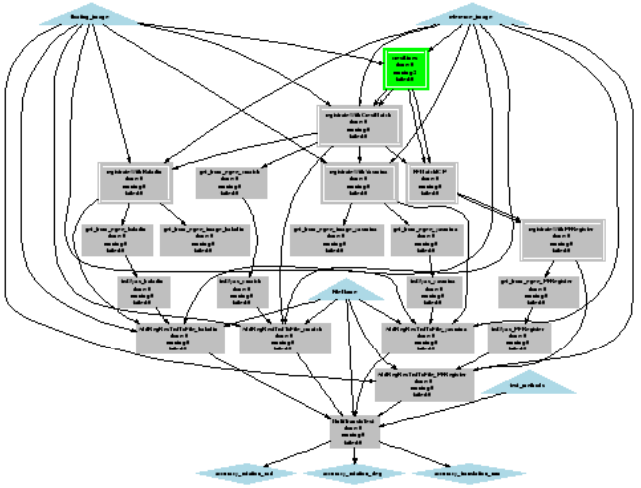




- **Characteristics**
  - Rapid response: a human waiting for the result!
  - Many small but CPU-intensive tasks
  - User is not aware of “grid”!
- **Needs**
  - Interfacing (data & computing) with non-grid application or portal
  - User and rights management between front-end and grid

Workflow

- **Examples**
  - "Bronze Standard": image registration
  - Flood prediction

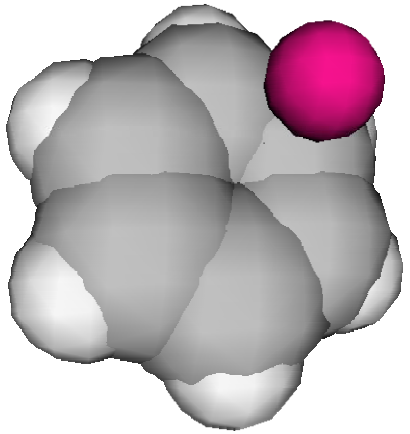


- **Characteristics**
  - Use of grid and non-grid services
  - Complex set of algorithms for the analysis
  - Complex dependencies between individual tasks

- **Needs**
  - Tools for managing the workflow itself
  - Standard interfaces for services (I.e. web-services)

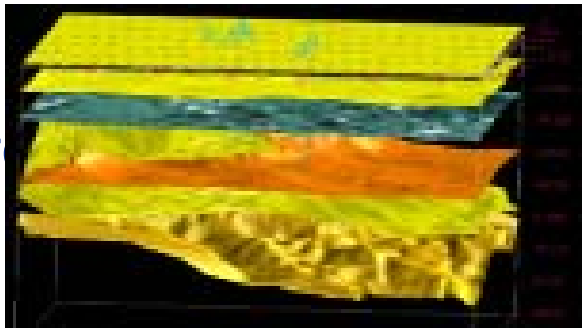
Parallel Jobs

- **Examples**
  - Climate modeling
  - Earthquake analysis
  - Computational chemistry
- **Characteristics**
  - Many interdependent, communicating tasks
  - Many CPUs needed simultaneously
  - Use of MPI libraries
- **Needs**
  - Configuration of resources for flexible use of MPI
  - Pre-installation of optimized MPI libraries



Legacy Applications

- **Examples**
  - Commercial or closed source binaries
  - Geocluster: geophysical analysis softwar
  - FlexX: molecular docking software
  - Matlab, Mathematics, …



- **Characteristics**
  - Licenses: control access to software on the grid
  - No recompilation ⇒ no direct use of grid APIs!
- **Needs**
  - License server and grid deployment model
  - Transparent access to data on the grid

Universal Needs

- **Security**
  - Ability to control access to services and to data
    - Fine-grained access control lists
    - Encryption & logging for more demanding disciplines
    - Access control consistently implemented over all services
- **VO Management**
  - Management of users, groups, and roles
  - Changing the priority of jobs for different users, groups, roles
  - Quota management for users, groups, roles
  - Definition and access to special resources
    - Application-level services
    - Responsive queues (guaranteed, low-latency execution)

Grid Middleware

- **Services exist for many of the application needs and plans exist to fix existing deficiencies or holes.**

- **No longer "one-size-fits-all" world:**
  - Works for low-level services (CPU, storage).
  - Higher-level services imply trade-offs:
    - E.g. latency vs. bulk response of meta-schedulers
    - E.g. security vs. speed for data access
  - Commonalities allow "one-size-fits-many" solutions.
- **Future evolution:**
  - Standards more important than ever: plug-and-play services.
  - Diversification of higher-level services is healthy and inevitable.
  - Integration of third-party tools an absolute necessity.

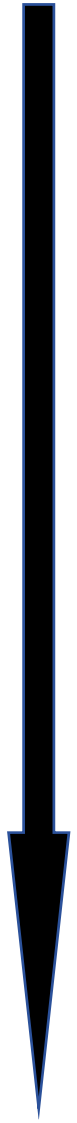
Overview

- **Two "pilot" application areas in EGEE**
  - High Energy Physics
  - Biomedical
- **Examples of the growing number of diverse communities**
- **What's being done?! A classification of applications**
- **Challenges for application developers**

- **NOTE**
  - Not giving details on each application but look for possible patterns and similarities with your own interests – perhaps in different disciplines
  - Grid middleware supports commonly met requirements so techniques will also cross between different communities

Different Goals for App. Development

- **I need resources for my research**
  - I need richer functionality
    - MPI, parametric sweeps,…
    - Data and compute services together…
- **I provide an application for (y)our research**
  - How!?
    - Pre-install executables ?
    - Hosting environment?
    - Share data
    - Use it via portal?
- ***We* provide applications for (y)our research**
  - Also need:
    - Coordination of development
    - Standards
    - …



Engineering challenges increasing

Challenges

- **Research software is often**
  - Created for one user: the developer
  - Familiarity makes it useable
  - Short-term goals: Used until papers are written and then discarded
- **Grid applications are often used**
  - by a VO
  - Without support from developer
  - In new contexts and workflows
- **Grid application developers are**
  - In a research environment
  - Yet their s/w must have:
    - Stability
    - Documentation
    - Useability
    - Extendability
  - i.e. Production quality

Need expertise in:

- **software engineering**
- **application domain**
- **grid computing**

Summary

- **Observe routine and large-scale use of the EGEE infrastructure by numerous, diverse set of users.**
- **EGEE provides backbone services which support wide range of different grid application families.**
  - Simulation, Bulk Processing, Responsive Apps., Workflow, Parallel Jobs, Legacy Applications
- **Third-party tools are becoming increasingly important for providing specialized (but flexible) services to particular groups of applications.**
- **NA4 website (http://egeena4.lal.in2p3.fr/)**

Participation

- **Related projects:**
  - DEGREE
  - DILIGENT
  - EGRID
  - EU ChinaGRID
  - EU MedGRID
  - GRIDCC
  - many more…
- **Other collaborations:**
  - Geant4
  - ITU
  - ProActive
  - many more…

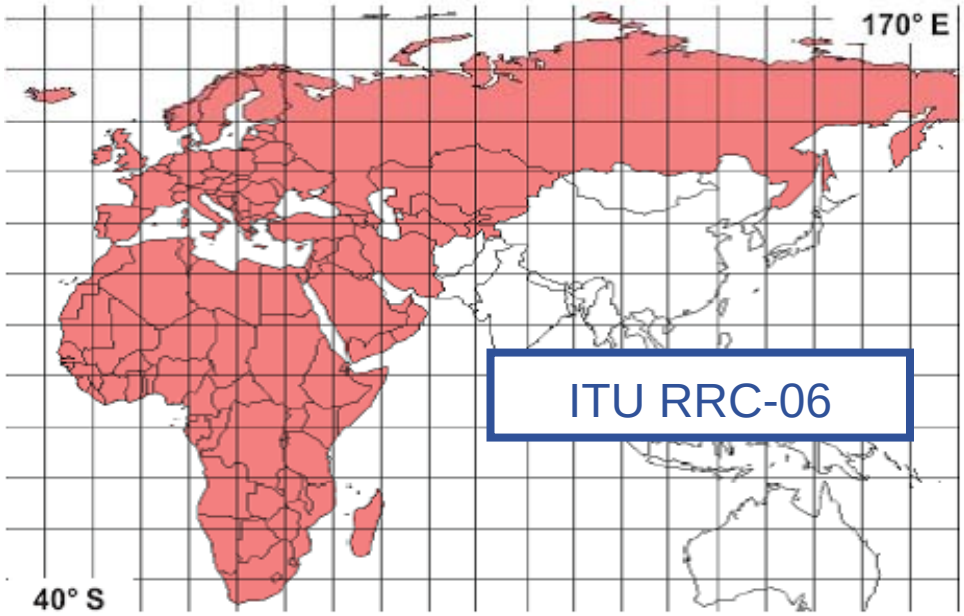


Figure 1
The extent of the planning area for the RRC-06

Participate

- **EGEE Conferences and Users' Forums**
  - Share your expertise, learn from other users.
  - Be open to collaboration with others.

- **Do (or don't) like something, speak up!**
  - VO issues, needs ⇒ VO Managers' Group
  - Resource, proc. problems ⇒ Operations Advisory Group (OAG)
  - Talk with NA4 steering committee
- **Report problems:**
  - Don't be afraid to use GGUS.
  - Report middleware annoyances ⇒ someone else is annoyed too!

NA4 website (http://egeena4.lal.in2p3.fr/)