The EGEE project – Building a Global Production Grid

Fabrizio Gagliardi
Project Director EGEE
CERN, Switzerland

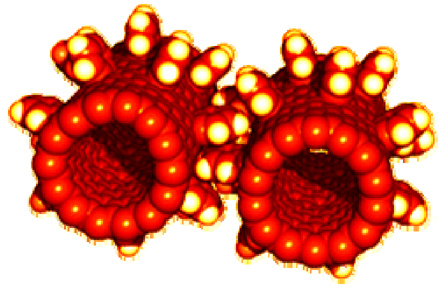
Vico Equense, Naples, 18 June 2005

The ISSGC

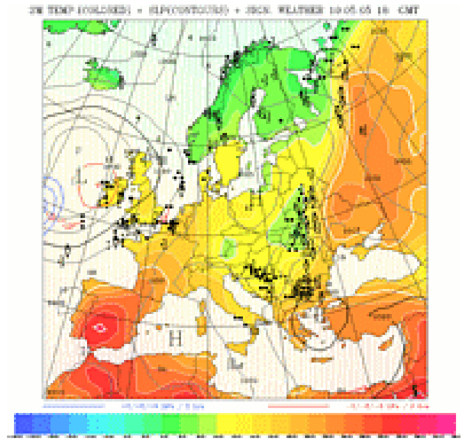
- **Grid computing is emerging as one of the most cost effective computing paradigms for a large class of data and compute intensive applications**
- **Still a a stiff entry cost for computational scientists**
- **CERN computing summer school initiated a Grid track back in 2002**
  - Extraordinary success
  - Idea to start a dedicated summer school on Grid computing
  - Following excellent response in 2003 and 2004, here we are with the third run in 2005
  - From next year will continue with EU supported ICEAGE

Computing intensive science

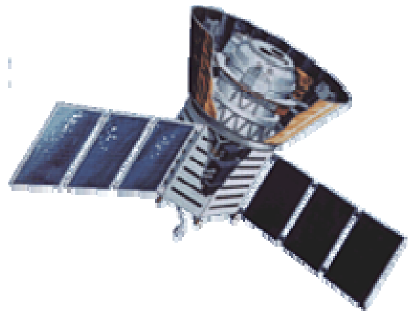
- **Science is becoming increasingly digital and needs to deal with increasing amounts of data**



- **Simulations get ever more detailed**
  - Nanotechnology – design of new materials from the molecular scale
  - Modelling and predicting complex systems (weather forecasting, river floods, earthquake)
  - Decoding the human genome



- **Experimental Science uses ever more sophisticated sensors to make precise measurements**
  - → Need high statistics
  - → Huge amounts of data
  - → Serves user communities around the world



The solution: the Grid

- **Integrating computing power and data storage capacities at major computer centres**
- **Providing users with seamless access to computing resources, 24/7, independent of geographic location**

→More effective and seamless collaboration of dispersed communities, both scientific and commercial

→Ability to run large-scale applications comprising thousands of computers, for wide range of applications

→The term “e-Science” has been coined to express these benefits

- **Objectives**
  - consistent, robust and secure service grid infrastructure
  - improving and maintaining the middleware
  - attracting new resources and users from industry as well as science
- **Structure**
  - 70 leading institutions in 27 countries, federated in regional Grids
  - leveraging national and regional grid activities worldwide
  - funded by the EU with ~32 M Euros for first 2 years starting 1st April 2004

EGEE Infrastructure

We are already exceeding original project plans

- (yellow marker) Country providing resources
- (red marker) Country anticipating joining EGEE/LCG

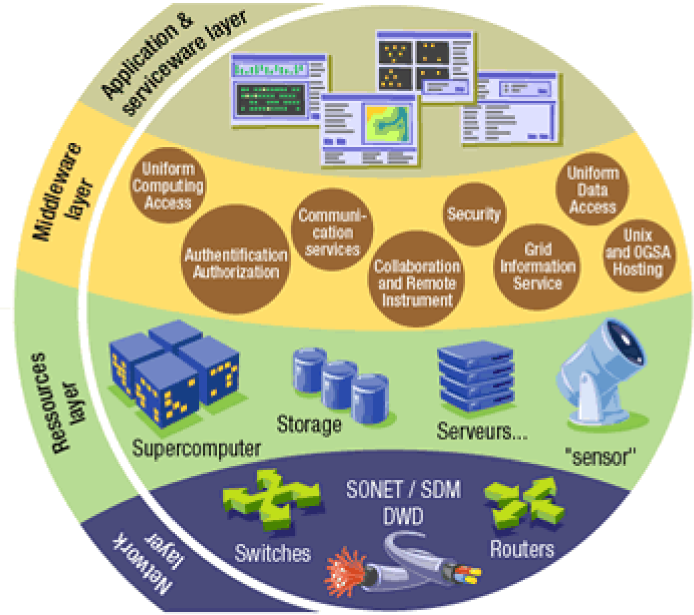
In EGEE-0:

- ⇨ >140 sites
- ⇨ >14 000 CPUs
- ⇨ >5 PB storage

Middleware

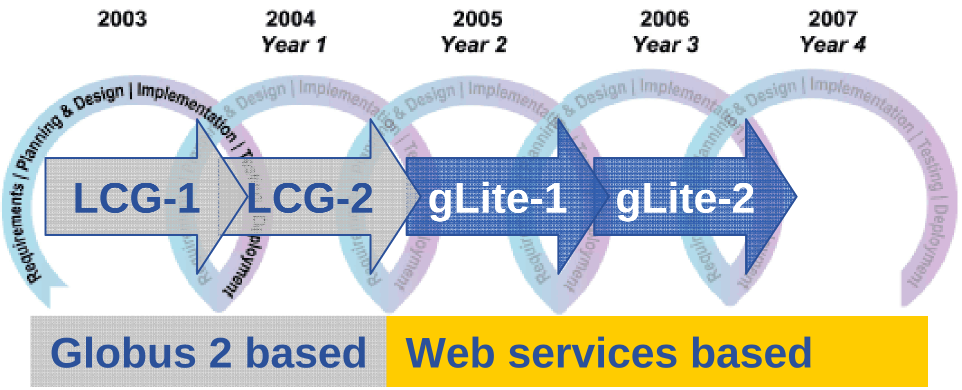
- **The Grid relies on advanced software, called middleware, which interfaces between resources and the applications**

- **The GRID middleware:**
  - Finds convenient places for the application to be run
  - Optimises use of resources
  - Organises efficient access to data
  - Deals with authentication to the different sites that are used
  - Runs the job & monitors progress
  - Recovers from problems
  - Transfers the result back to the scientist



EGEE Middleware - gLite

- Intended to replace present middleware with production quality services
- Developed from existing components
- Aims to address present shortcomings and advanced needs from applications
- Prototyping short development cycles for fast user feedback
- Initial web-services based prototypes being tested



Application requirements http://egee-na4.ct.infn.it/requirements/

gLite

- **First release of gLite end of March 2005**
  - Focus on providing users early access to prototype
- **Lightweight services**
- **Interoperability & Co-existence with deployed infrastructure**
- **Robust: Performance & Fault Tolerance**
- **Portable**
- **Service oriented approach**
  - Follow WSRF standardisation
- **Site autonomy**
- **Open source license**

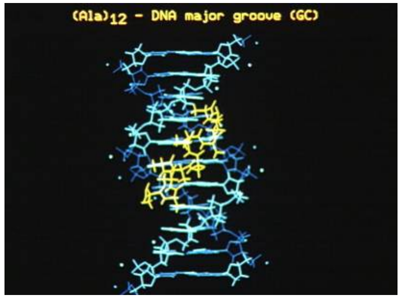
gLite Interoperability

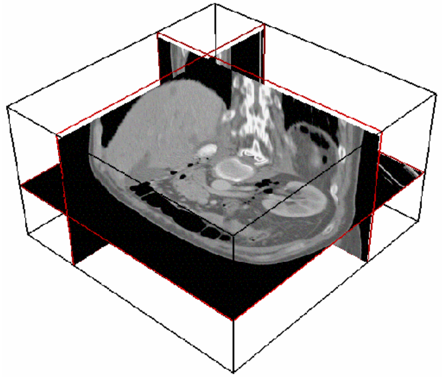
- **Interoperability with other Grids mainly needed at resource level**
  - Same physical resource should be exploitable in different Grids
- **Approach**
  - Reduce requirements on sites
    - Computing Element: globus gatekeeper
    - Storage Element: SRM
  - Close connection with other projects
    - OSG
      - *Use EGEE architecture and design documents as basis for their blueprint*
      - *Common members in design teams*

Pilot applications

- **High-Energy Physics (HEP)**
  - Provides computing infrastructure (LCG)
  - Challenging:
    - thousands of processors world-wide
    - generating terabytes of data
    - ‘chaotic’ use of grid with individual user analysis (thousands of users interactively operating within experiment VOs)

- **Biomedical Applications**
  - Similar computing and data storage requirements
  - Major challenge: security



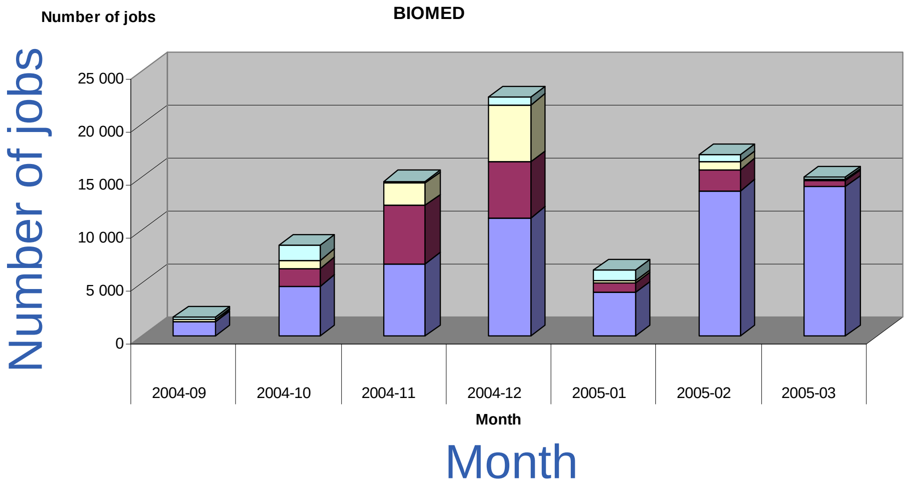


BioMed Overview

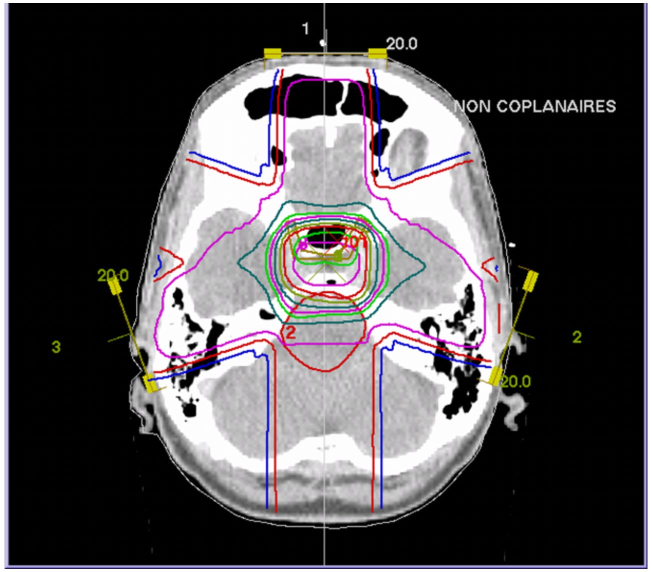
- **Infrastructure**
  - ~2.000 CPUs
  - ~21 TB disks
  - in 12 countries

- **>50 users in 7 countries working with 12 applications**
- **18 research labs**

- **~80.000 jobs launched since 04/2004**
- **~10 CPU years**

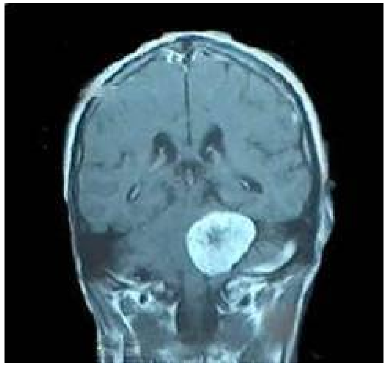


- **GEANT4 Application to Tomography Emission**
  - Scientific objectives
    - Radiotherapy planning to improve treatment of tumors computed from pre-treatment MR scans
  - Method
    - GEANT4-based software to model physics of nuclear medicine
    - Monte Carlo simulation to improve accuracy of computations



  - Grid added value
    - Splitting the random number sequences needed for Monte Carlo simulations enables independent computations
    - Parallelization reduces the total computation time
  - Results and perspectives
    - computation time reduced BUT not sufficiently for clinical practice
      → further optimizations are on-going
    - large community of users is interested in GATE

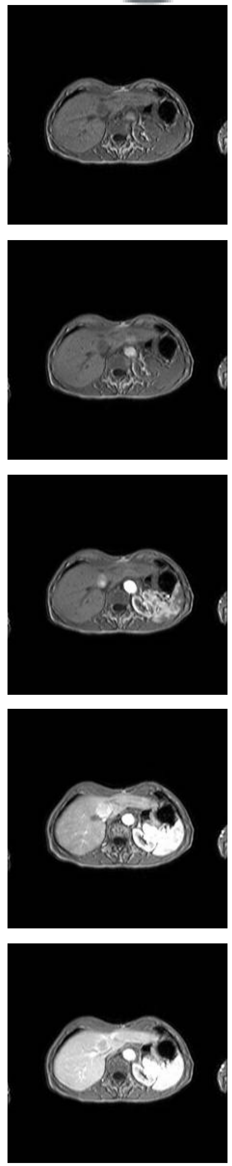
- **Clinical Decision Support System**
  - Scientific objectives
    - Extract clinically relevant knowledge to guide practitioners in their clinical practice
  - Method
    - Starting from trained databases
    - Use classifier engines
    - Compare to annotated databases to classify data
  - Grid added value
    - Ubiquitous access to distributed databases and classifier engines
    - Grid information system to publish and discover data sources and engines
    - Automatic management of login and security
  - Results and perspectives
    - 12 classification engines available
    - 1000 medical cases registered
    - Dynamic discovery of all engines can be implemented on top of the grid information system
    - Accounting will be provided by the grid



Classification of tumours in soft tissues

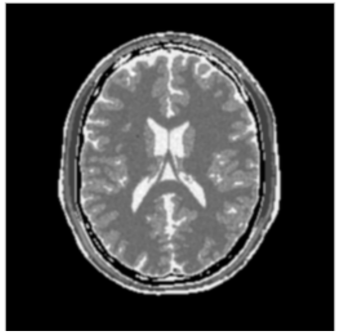
Pharmacokinetics

- **Co-registration of Medical Images**
  - Scientific objectives
    - Contrast Agent Diffusion to characterize tumour tissues without biopsy
  - Method
    - Co-registration requires deformable registration methods
      → compute intensive
  - Grid added value
    - Processing of compute intensive co-registration and generation of diffusion maps for the 3D MRI Studies.
    - Parallel & independent computations on different input data sets
  - Results and perspectives
    - Last clinical test:
      12 patients with 13 MRI studies each
      each study comprises 24 512x512 12-bit slices
    - Processing of the registration algorithm takes around 12 hours per study
    - Registration parameters tuned with four possible combinations
    - Each combination of parameter took 2 hours
      → 72 times faster than with a single computer

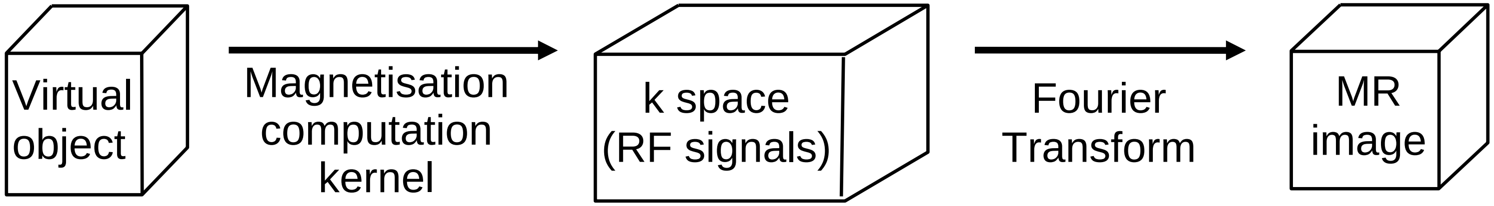


Grid Protein Structure Analysis

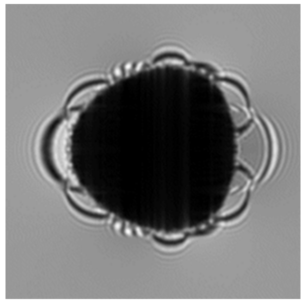
- Scientific objectives
  - Integrating up-to-date databases and relevant algorithms for bio-informatic analysis of data from genome sequencing projects
- Method
  - Protein databases are stored on the grid as flat files
  - Protein sequence analysis tools run unchanged on grid resources
  - Output is analysed and displayed in graphic format through the web interface
- Grid added value
  - Convenient way to distribute and access international databanks, and to store more and larger databases
  - Compute larger datasets with available algorithms
  - Open to a wider user community
- Results and perspectives
  - 9 bioinformatic softwares gridified so far
  - large number of rather short jobs (few minutes each)
  - Optimizations on-going to
    - *speed up access to databases*
    - *lower short jobs latencies*



- **3D Magnetic Resonance Image Simulator**
  - Scientific objectives
    - Better understand MR physics by studying MR sequences *in silico* and MR artefacts
    - Validate MR Image processing algorithms on synthetic but realistic images
  - Method

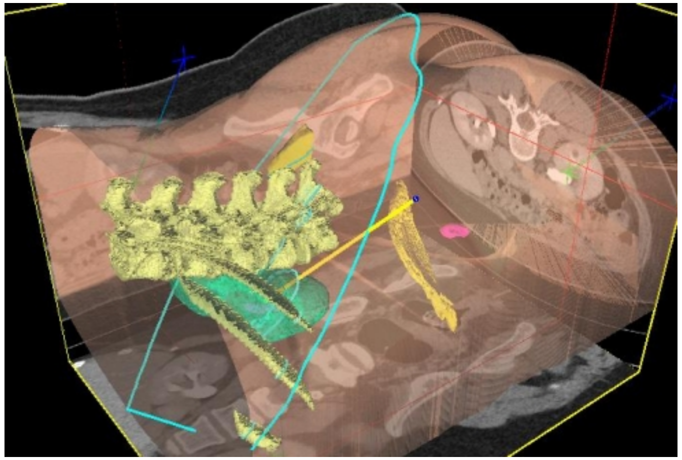


  - Grid added value
    - Speeds up the simulation time
    - Enables simulation of high resolution images
    - Offers an access to MPI-enabled clusters
  - Results and perspectives
    - Manageable computation time for medium size images
    - Development of a portal to ease access to the application
    - Implementation of new artifacts



gPTM3D

- **3D Medical Image Analysis Software**
  - Scientific objectives
    - Interactive volume reconstruction on large radiological data
  - Method
    - Starting from hand-made initialization
    - Algorithm segments each slice of a medical volume
    - 3D reconstruction by triangulating contours from consecutive slices
  - Grid added value
    - Interactive reconstruction time: less than 2mins and scalable
    - Permanent availability of resources for fast reconstruction
    - Access to users at non grid-enabled sites (e.g. hospital)
    - Unmodified medically optimized interface
  - Results and perspectives
    - Successfully ported and demonstrated at first EGEE review
    - Streams to/from non EGEE-enabled sites specific protocol, CrossGrid glogin will be considered
    - Resource access QoS: ongoing work



xmipp_ML_refine

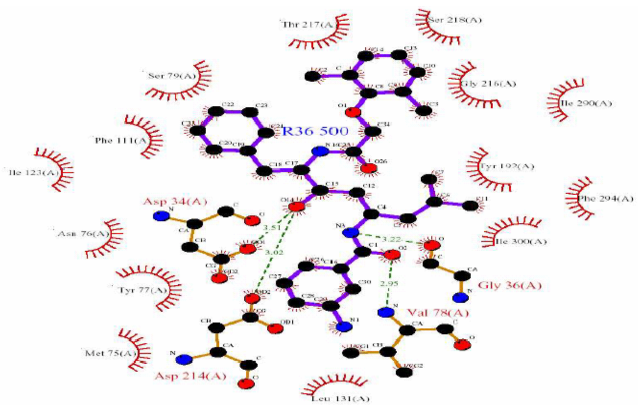
- **Macromolecules structure analysis from electron microscopy**
  - Scientific objectives
    - 3D reconstruction of molecular structural information from cryo-electron microscopy
  - Method
    - Multi-reference refinement of electron microscopy structures through a maximum likelihood statistical approach
  - Grid added value
    - Very compute intensive analysis of multiple structures
      - *2D: one to several weeks on a single CPU*
      - *3D: even more costly*
    - Computation can be split in independent jobs that are executed in parallel
  - Results and perspectives
    - First results on 2D analysis show significant time gain:
      two months on a local cluster (20 CPUs) versus one month on the grid
    - algorithm still being optimized and ported to 3D case
    - MPI implementation is currently being developed that should significantly improve the computation time

xmipp_ML_CTFs

- **Electron microscope images correction**
  - Scientific objectives
    - Electron microscopy images impaired by electron sources and defocus of magnetic lenses used in experimental practice
    - Image aberrations are described by a Contrast Transfer Function (CTF) that need to be estimated to fix images
    - CTF estimation lead to drastic image enhancement
  - Method
    - Auto regressive modelling is used to estimate parameters of the CTF and produce more reliable results than classical Fourier transform-based approaches
  - Grid added value
    - Very compute intensive: complex functional, slow optimisation process
    - Parallelisation on different grid resources
  - Results and perspectives
    - → 2 months on a single CPU
    - → 2 days on a local 20-CPUs cluster
    - → 14 hours on the grid

In silico Drug Discovery

– Scientific objectives
  - Provide docking information to help in the search for new drugs
  - Propose new inhibitors (drug candidates) addressed to neglected diseases
  - *In silico* virtual screening of drug candidate databases
– Method
  - Large scale molecular docking on malaria to compute millions of potential drugs with different software and parameters settings
– Grid added value
  - Drug discovery usually takes up to 12 years to complete
  - Docking much faster, but large databases lead to heavy computations
    → split candidate drug input on different grid resources
– Results and perspectives
  - Limited size computation (105 candidate drugs tested for 1 protein target) achievable in 2 days using the Grid compared to 6 months of CPU time
  - Full data challenge planed
    - *3x106 candidate drugs against 5 protein targets*
    - *Total computing time will reach 80 years of CPU and 6 TB of storage*



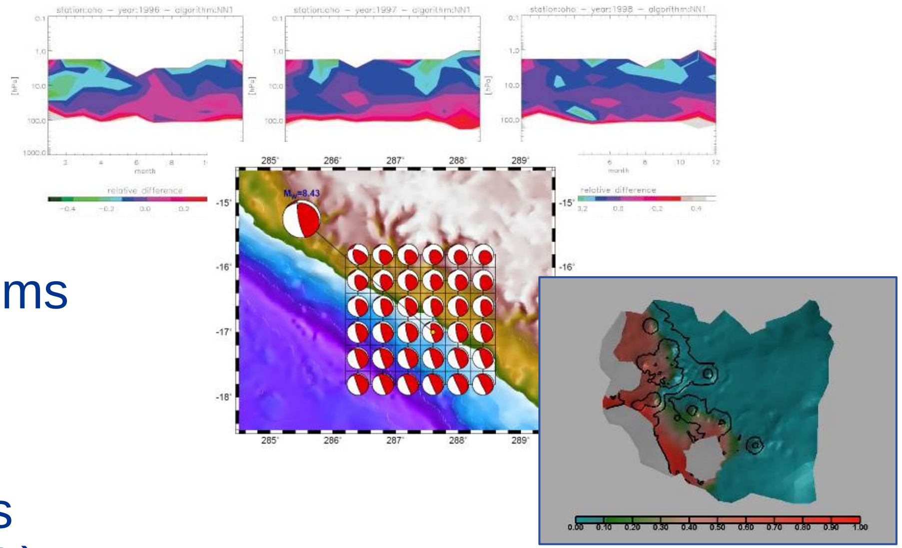
SPLATCHE

- **Genome evolution modeling**
  - Scientific objectives
    - Study human evolutionary genetics and answer questions such as
      - *geographic origin of modern human populations*
      - *genetic signature of expanding populations*
      - *genetic contacts between modern humans and Neanderthals*
  - Method
    - Simulate past demography of human populations in a geographically realistic landscape
    - Generate molecular diversity of samples of genes drawn from the current human's range, and compare to observed contemporary molecular diversity
  - Grid added value
    - Due to the Bayesian approach used, the SPLATCHE application is very compute intensive
    - Independent simulations can be executed in parallel
  - Results and perspectives
    - Application prototype ported on the EGEE middleware
    - Scale tests on the full grid infrastructure underway

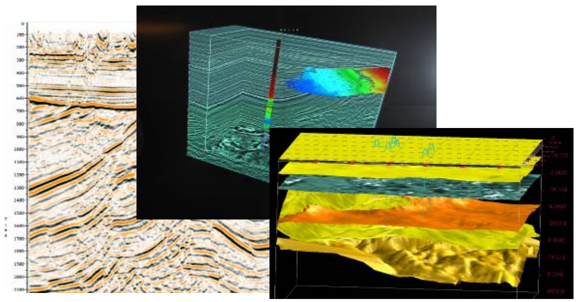
Generic Applications

- **EGEE Generic Applications Advisory Panel (EGAAP)**
  - UNIQUE entry point for “external” applications
  - Reviews proposals and make recommendations to EGEE management
    - Deals with “scientific” aspects, not with technical details
    - Generic Applications group in charge of introducing selected applications to the EGEE infrastructure
  - 6 applications selected so far:
    - Earth sciences (I and II)
    - MAGIC
    - Computational Chemistry
    - PLANCK
    - Drug Discovery
    - GRACE (end Feb 2005)

Earth sciences applications



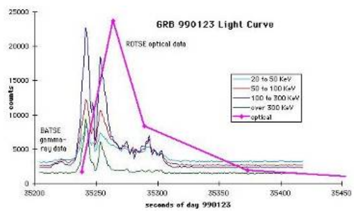
- **Earth Observations by Satellite**
  - ozone profiles
- **Solid Earth Physics**
  - Fast Determination of mechanisms of important earthquakes
- **Hydrology**
  - Management of water resources in Mediterranean area (SWIMED)
- **Geology**
  - Geocluster: R&D initiative of the Compagnie Générale de Géophysique

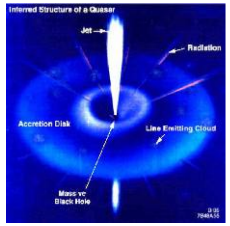


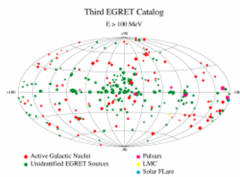
- **A large variety of applications ported on EGEE which incites new users**
- **Interactive Collaboration of the teams around a project**

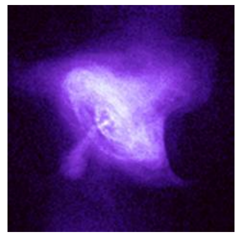
MAGIC

- **Ground based Air Cerenkov Telescope 17 m diameter**
- **Physics Goals:**
  - Origin of VHE Gamma rays
  - Active Galactic Nuclei
  - Supernova Remnants
  - Unidentified EGRET sources
  - Gamma Ray Burst
- **MAGIC II will come 2007**
- **Grid added value**
  - Enable “(e-)scientific“ collaboration between partners
  - Enable the cooperation between different experiments
  - Enable the participation on Virtual Observatories

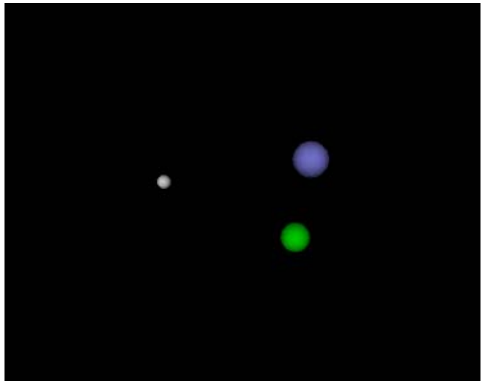


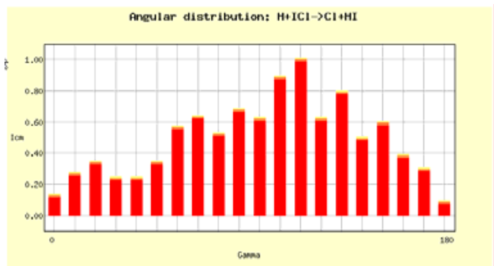


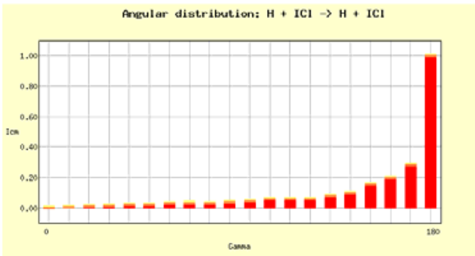


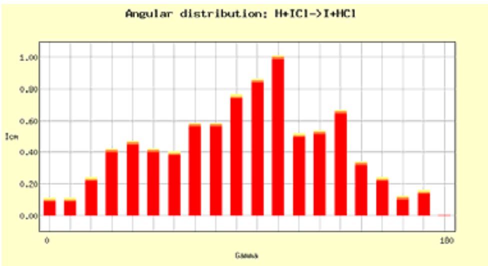


- **The Grid Enabled Molecular Simulator (GEMS)**
  - Motivation:
    - Modern computer simulations of biomolecular systems produce an abundance of data, which could be reused several times by different researchers.
      → data must be catalogued and searchable
  - GEMS database and toolkit:
    - autonomous storage resources
    - metadata specification
    - automatic storage allocation and replication policies
    - interface for distributed computation







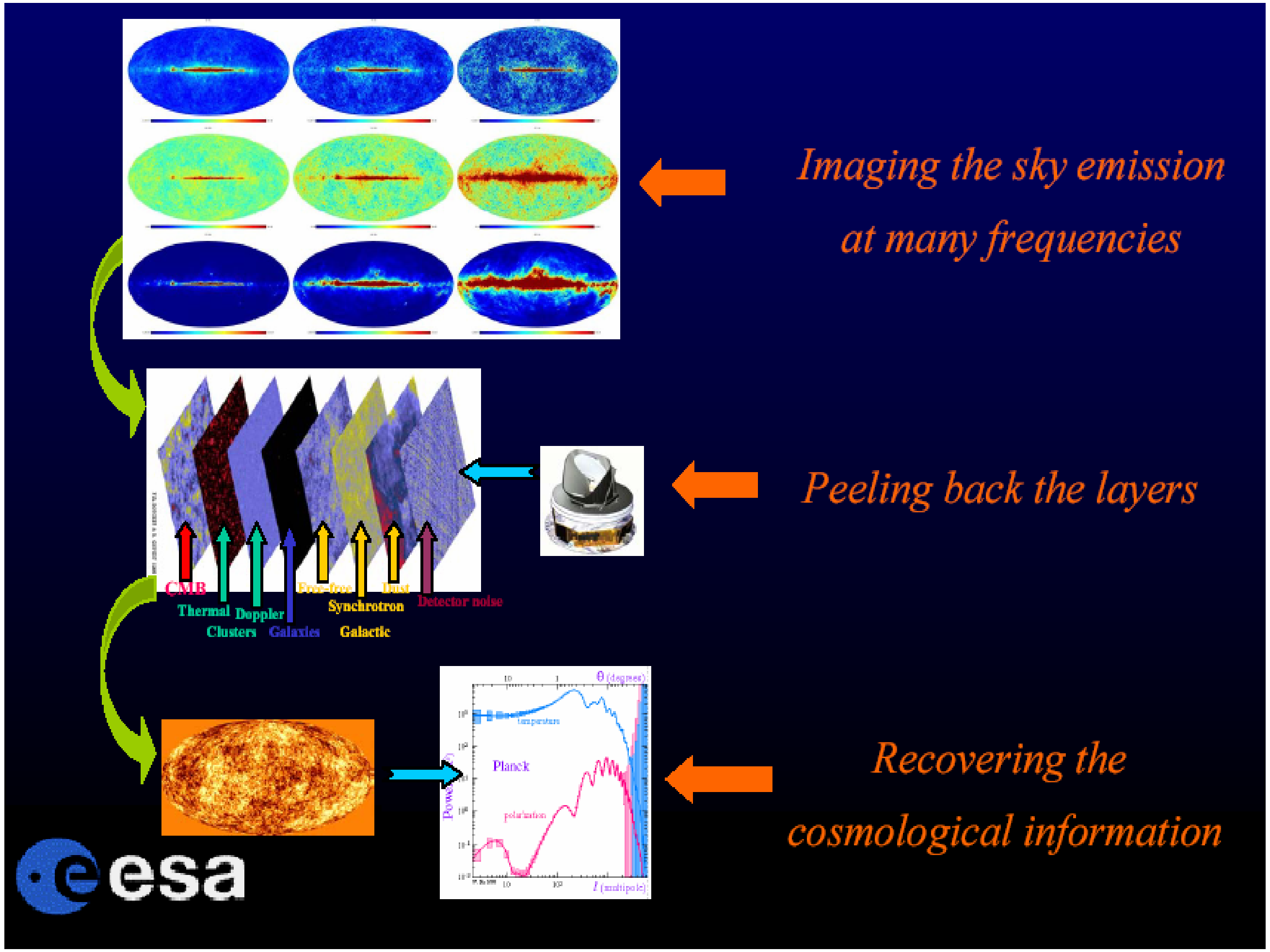


Planck

- **On the Grid:**
  > 12 time faster
  (but ~5% failures)

- **Complex data structure**
  → data handling important

- **The Grid as**
  - collaboration tool
  - common user-interface
  - flexible environment
  - new approach to data and S/W sharing



User information & support

- **More than 140 training events (including the GGF grid school) across many countries**
  - >1200 people trained
    induction; application developer; advanced; retreats
  - Material archive coming online with ~200 presentations

- **Public and technical websites constantly evolving to expand information available and keep it up to date**

- **3 conferences organized**
  ~ 300 @ Cork
  ~ 400 @ Den Haag
  ~450 @ Athens

- **Pisa: 4th project conference 24-28 October ’05**

Collaborations

- **EGEE closely collaborates with other projects, e.g.**

- **Flooding Crisis (CrossGrid) demonstrated at 3rd EGEE conference in Athens**
  - Simulation of flooding scenarios
  - Display in Virtual Reality
  - Optimize data transport

→ **won prize for "best demo"**

Collaboration with Slowak Academy of Sciences

EGEE as partner

- **Ongoing collaborations**
  - with non EU partners in EGEE: US, Israel, Russia, Korea, Taiwan…
  - with other European projects, in particular:
    - GÉANT
    - DEISA
    - SEE-GRID
  - with non-European projects:
    - OSG: OpenScienceGrid (USA)
    - NAREGI
- **EGEE as incubator**
  - 16 recently submitted EU proposals supported, among them:
    - Baltic states (Baltic Grid proposal to EU)
    - Latin America (EELA consortium on ALIS/CLARA networking)
    - Mediterranean Area (EUMedConnect)
    - China: EUGridChina
- **EGEE supports Euro-India ICT Co-operation Initiative**

Related projects under negotiation

Name	*Description*	*Common partners with EGEE*
BalticGrid	EGEE extension to Estonia, Latvia, Lithuania	KTH – PSNC – CERN
EELA	EGEE extension to Brazil, Chile, Cuba, Mexico, Argentina	CSIC – UPV – INFN – CERN – LIP – RED.ES
EUChinaGRID	EGEE extension to China	INFN – CERN – DANTE – GARR – GRNET – IHEP
EUMedGRID	EGEE extension to Malta, Algeria, Morocco, Egypt, Syria, Tunisia, Turkey	INFN – CERN – DANTE – GARR – GRNET – RED.ES
ISSeG	Site security	CERN – CSSI – FZK – CCLRC
eIRGSP	Policies	CERN – GRNET
ETICS	Repository, Testing	CERN – INFN – UWM
ICEAGE	Repository for Training & Education, Schools on Grid Computing	UEDIN – CERN – KTH – SZTAKI
BELIEF	Digital Library of Grid documentation, organisation of workshops, conferences	UWM
BIOINFOGRID	Biomedical	INFN – CNRS
Health-e-Child	Biomedical – Integration of heterogeneous biomedical information for improved healthcare	CERN

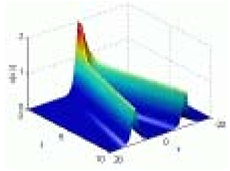
Exact budget and partner roles to be confirmed during negotiation

EGEE and Industry

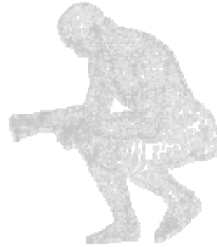
- **Industry as**
  - partner – to increase know-how on Grid technologies
  - user – for R&D applications
  - provider – of established Grid services, such as call centres, support centres and computing resource provider centres

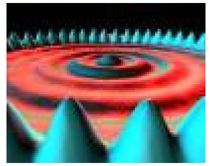
- **Industry Forum**
  - Raise awareness of the project among industries
  - Encourage businesses to participate
    - → ability to “experience” EGEE Grid in early stages

From Phase I to II

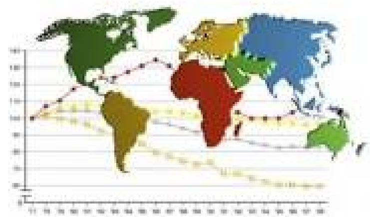


- **From 1st EGEE EU Review in February 2005:**
  - “The reviewers found the overall performance of the project very good.”
  - “… remarkable achievement to set up this consortium, to realize appropriate structures to provide the necessary leadership, and to cope with changing requirements.”

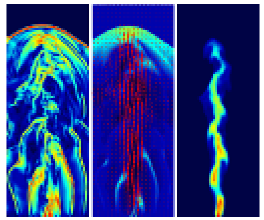


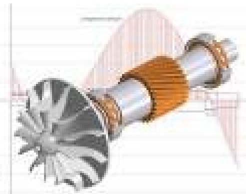


- **EGEE I**
  - Large scale deployment of EGEE infrastructure to deliver production level Grid services with selected number of applications



- **EGEE II**
  - Natural continuation of the project’s first phase
  - Emphasis on providing an infrastructure for e-Science
    - → increased support for applications
    - → increased multidisciplinary Grid infrastructure
    - → more involvement from Industry
  - Extending the Grid infrastructure world-wide
    - → increased international collaboration





Conclusions I

- Grid deployment are creating a powerful new tool for science – as well as applications from other fields
- Several applications are already benefiting from Grid technologies
- Investments in grid projects are growing world-wide
- Europe is strong in the development of Grids also thanks to the success of EGEE

Conclusions II

- **Collaboration across national and international programmes is very important:**
  - Grids are above all about collaboration at a large scale
  - Science is international and therefore requires an international computing infrastructure
- **ISSGC is now inspiring other similar schools such as the regional EGEE school in Budapest, other schools in the US and elsewhere**
- **Scientists attending these schools promote adoption of Grid computing worldwide**
- **Important to receive feedback and inputs from the school participants**
- **EGEE will be presented in more details during the rest of this week**

Contacts

- **EGEE Website**

  http://www.eu-egee.org

- **How to join**

  http://public.eu-egee.org/join/

- **EGEE Project Office**

  project-eu-egee-po@cern.ch