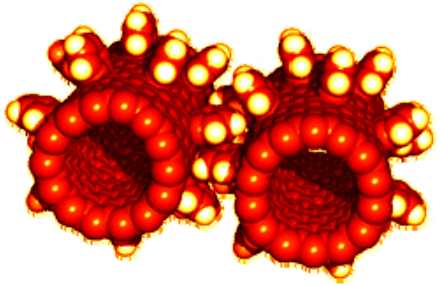
# Introduction to EGEE

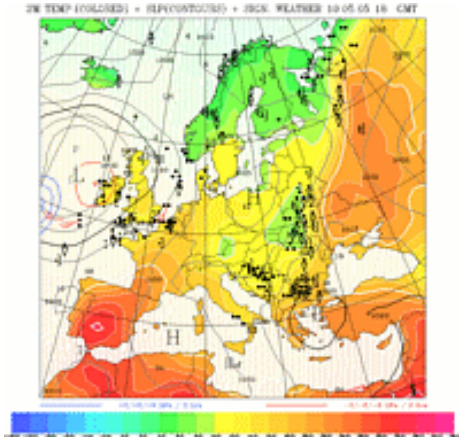
*Roberto Barbera*
*University of Catania and INFN*
*EGEE Tutorial*
*Roma, 02.11.2005*

# Computing intensive science

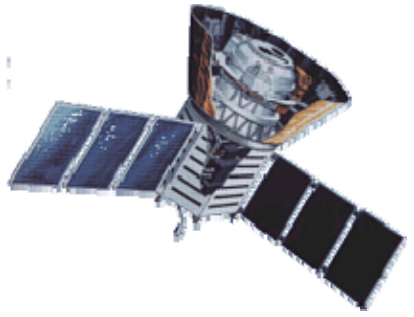
- **Science is becoming increasingly digital and needs to deal with increasing amounts of data**



- **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

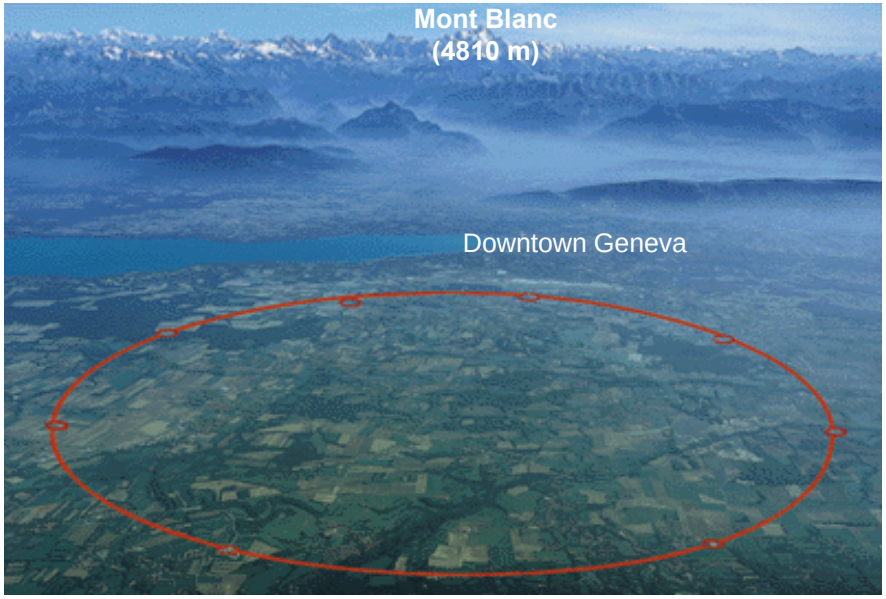


- **Experimental Science uses ever more sophisticated sensors to make precise measurements**
  - → Need high statistics
  - → Huge amounts of data
  - → Serves user communities around the world



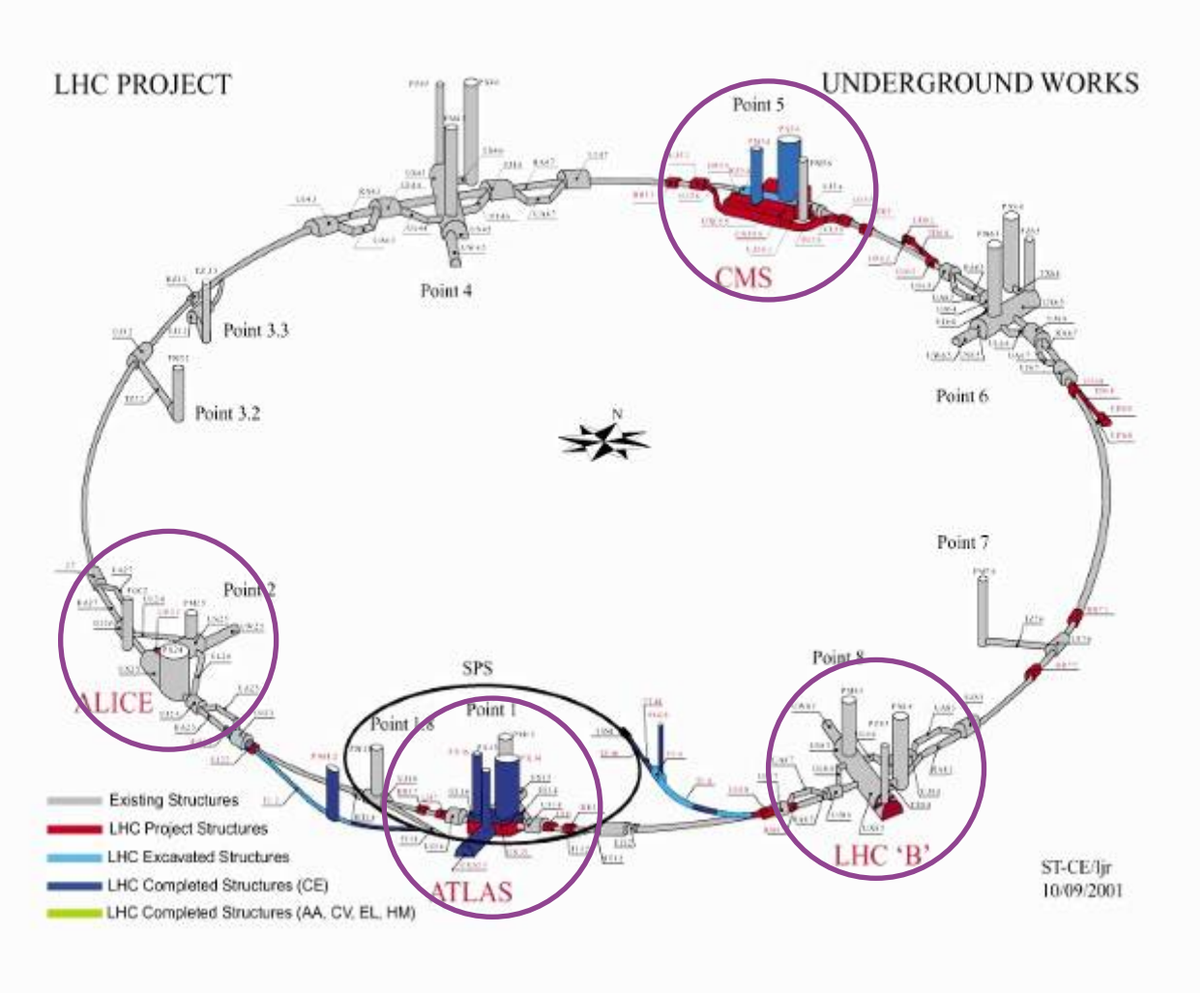
# A good example: Particle Physics

- Large amount of data produced in a few places: CERN, FNAL, KEK…
- Large worldwide organized collaborations (i.e. LHC CERN experiments) of computer-savvy scientists
- Computing and data management resources distributed world-wide owned and managed by many different entities
- **Large Hadron Collider (LHC) at CERN in Geneva Switzerland:**
  - One of the most powerful instruments ever built to investigate matter



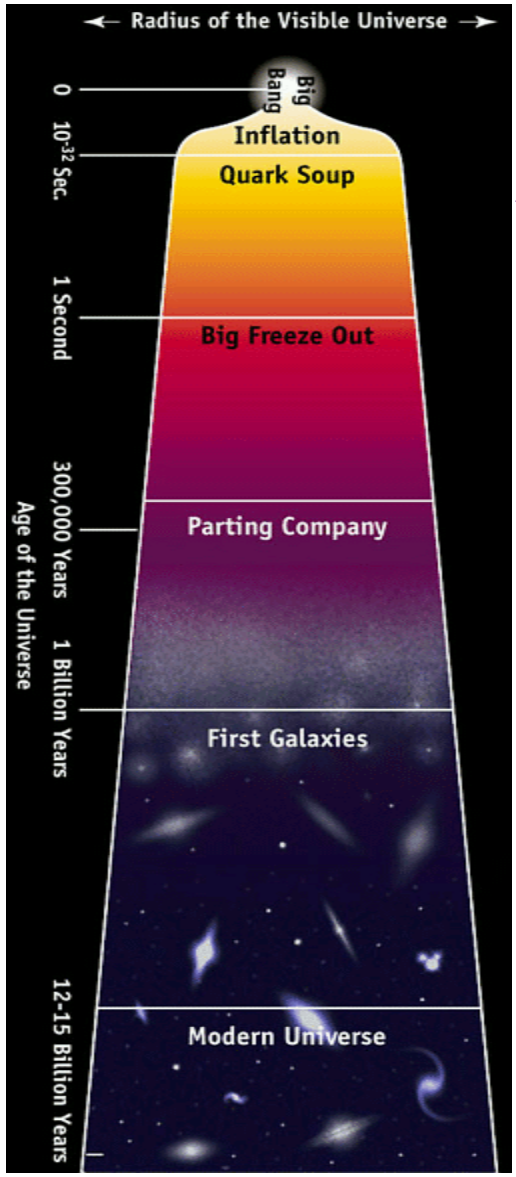
# The LHC Experiments

- **Large Hadron Collider (LHC):**
  - four experiments:
    - ALICE
    - ATLAS
    - CMS
    - LHCb
  - 27 km tunnel
  - Start-up in 2007

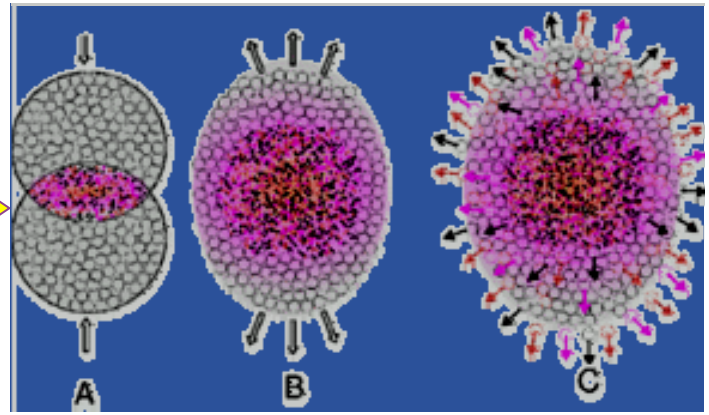


# ALICE goal: the search for the Quark Gluon Plasma

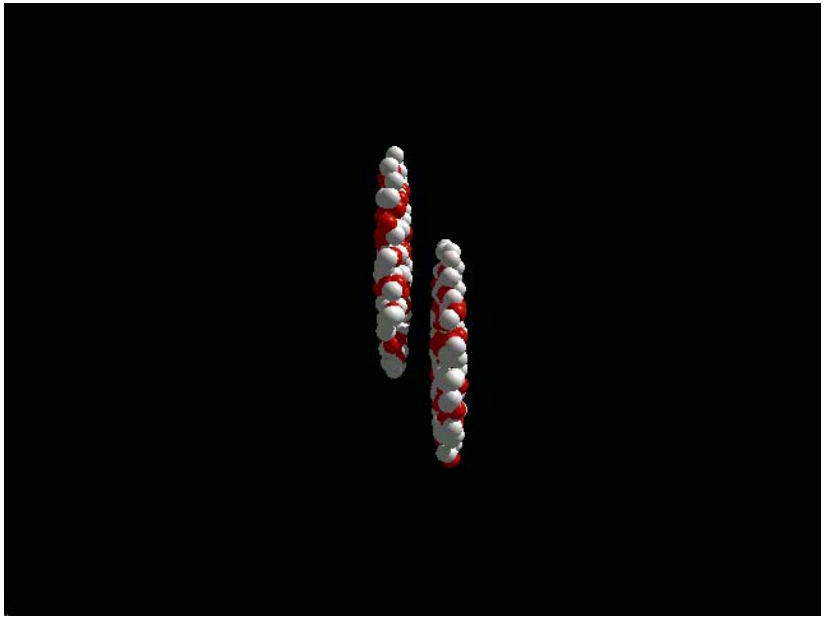
*The Big Bang*




*The Little Bang*

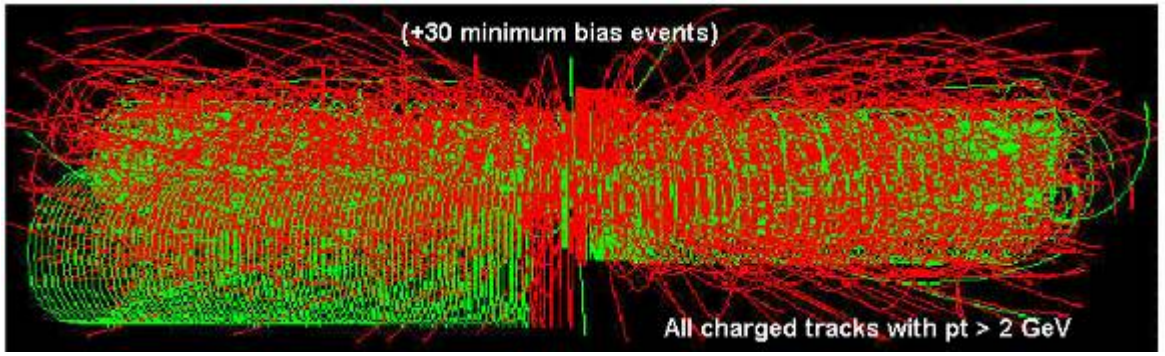



*Pb+Pb @ LHC (5.5 A TeV)*

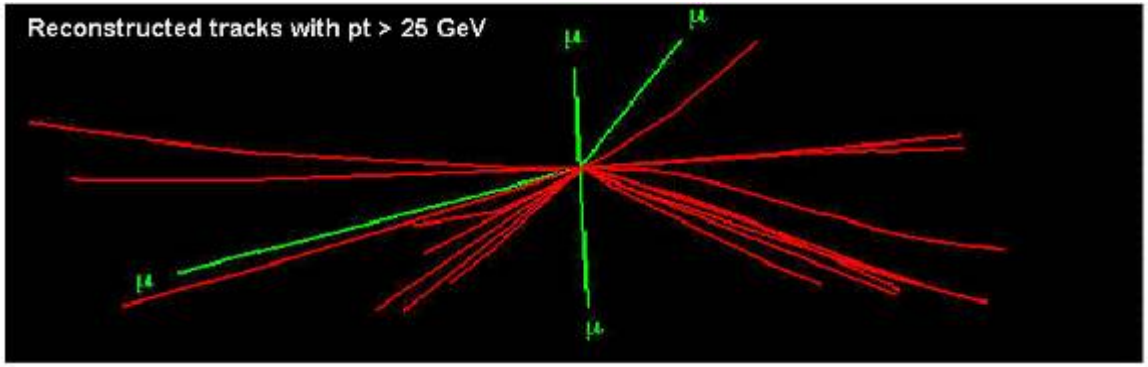


# The LHC Data Challenge

**Starting from this event**




**Looking for this "signature"**



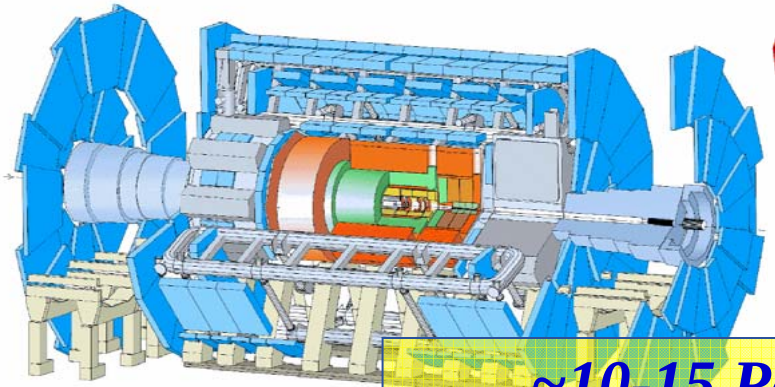

→ **Selectivity: 1 in $10^{13}$**

(Like looking for a needle in 20 million haystacks)

# The LHC Experiments

ATLAS

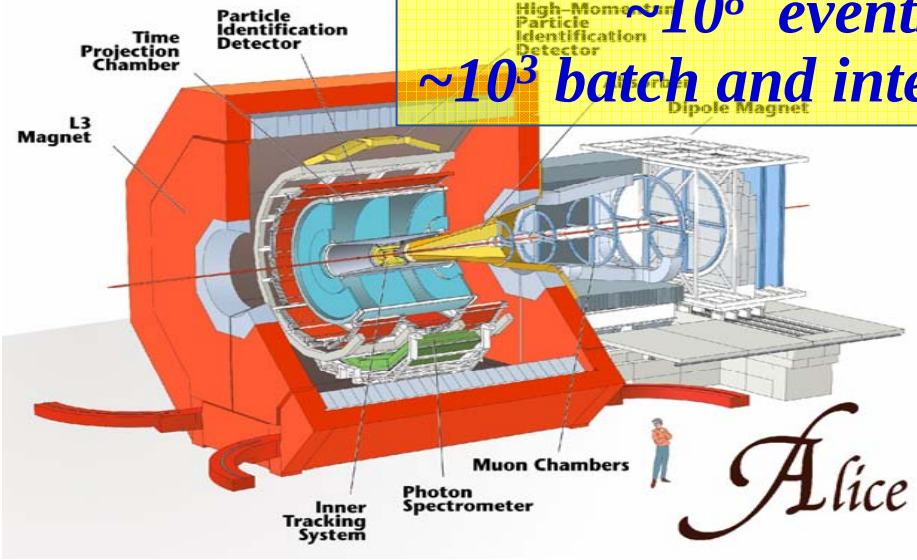
CMS

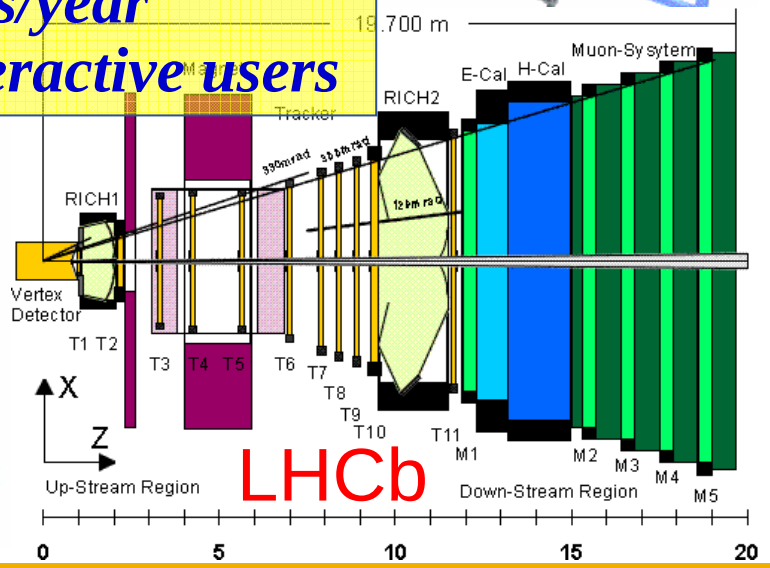


***~10-15 PetaBytes /year***

***~$10^8$ events/year***

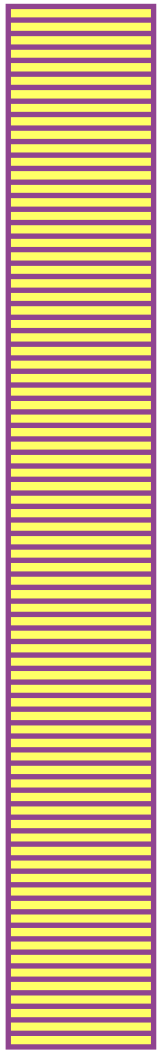
***~$10^3$ batch and interactive users***

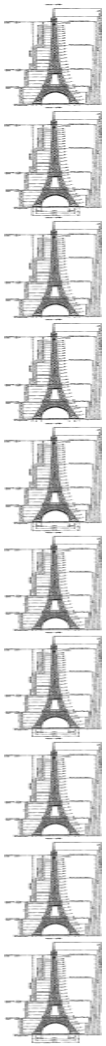


# Orders of magnitude…

10-15 Petabytes
~20.000.000 CD-ROM





10 times the Eiffel Tower
~3000 m

# The solution: the Grid

- Integrating computing and storage capacities at major computer centres
- 24/7 access, independent of geographic location

→ Effective and seamless collaboration of dispersed communities, both scientific and commercial

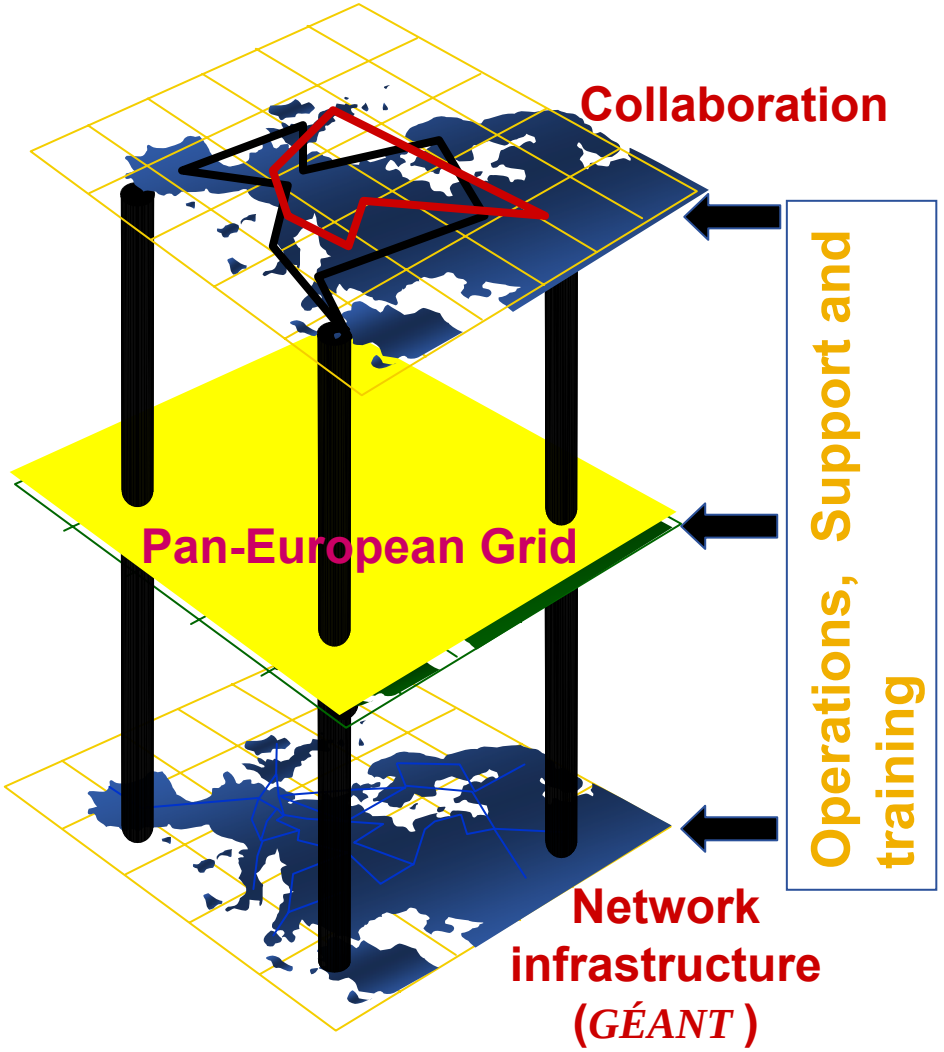
→ Ability to use thousands of computers for a wide range of applications

- Best cost effective solution for HEP LHC Computing Grid project (LCG) and from this the close integration of LCG and EGEE projects

# EGEE (Enabling Grids for E-sciencE)

Build a large-scale production grid service to:

- **Underpin European science and technology**
- **Link with and build on national, regional and international initiatives**
- **Foster international cooperation both in the creation and the use of the e-infrastructure**



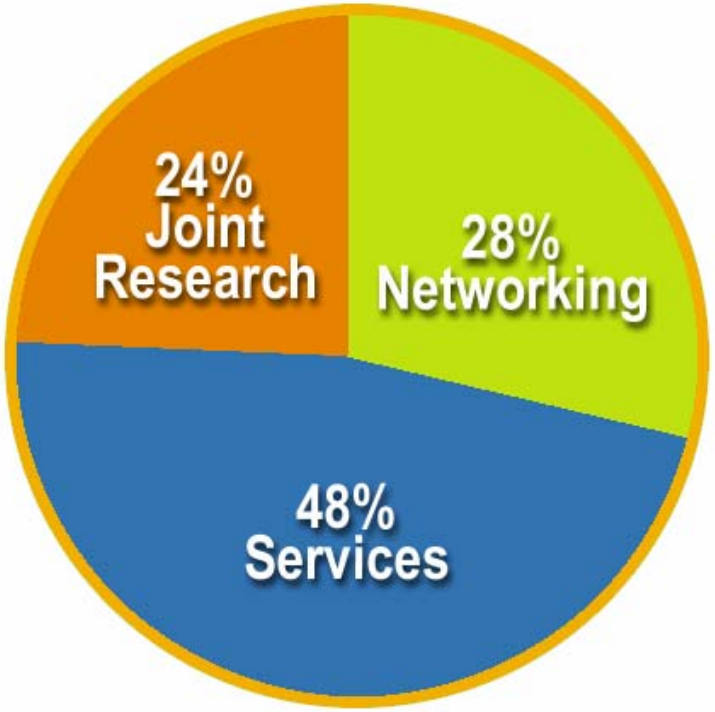
# The largest e-Infrastructure: EGEE

- **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**
  - 71 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

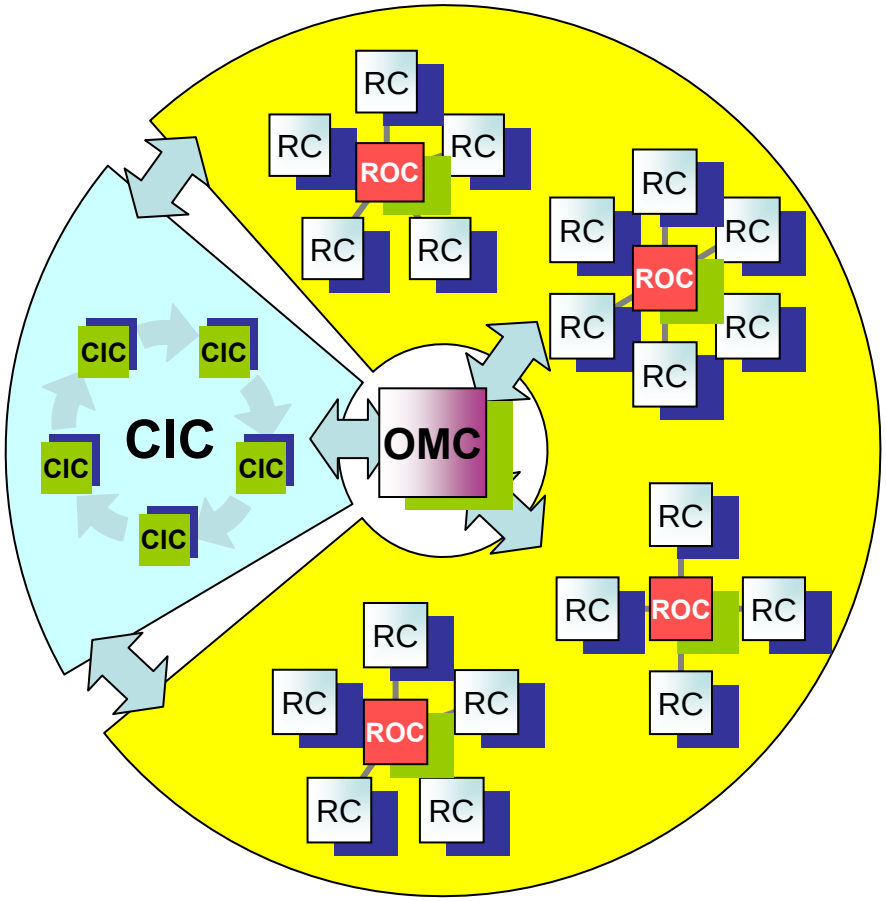
# EGEE Activities

- 48 % service activities (Grid Operations, Support and Management, Network Resource Provision)
- 24 % middleware re-engineering (Quality Assurance, Security, Network Services Development)
- 28 % networking (Management, Dissemination and Outreach, User Training and Education, Application Identification and Support, Policy and International Cooperation)




**Emphasis in EGEE is on operating a production grid and supporting the end-users**

# Grid Operations

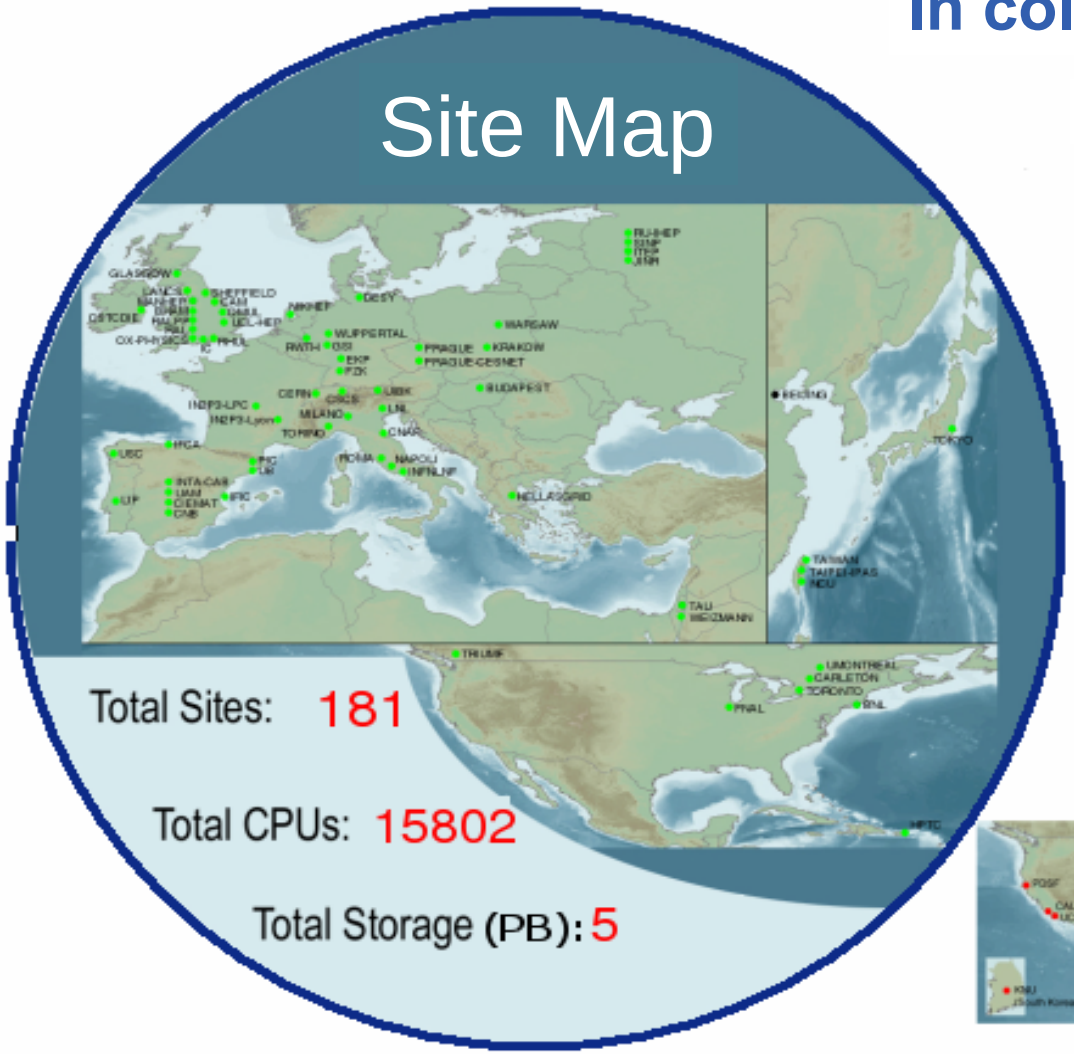



RC = Resource Centre
ROC = Regional Operations Centre
CIC = Core Infrastructure Centre
OMC = Operations Management Centre

- **The *grid* is flat, but**
- ***Hierarchy* of responsibility**
  - Essential to scale the operation
- **CICs act as a single Operations Centre**
  - Operational oversight ***(grid operator)*** responsibility
  - rotates weekly between CICs
  - Report problems to ROC/RC
  - ROC is *responsible* for ensuring problem is resolved
  - ROC oversees regional RCs
- **ROCs responsible for organising the operations in a region**
  - Coordinate deployment of middleware, etc
- **CERN coordinates sites not associated with a ROC**

# EGEE Infrastructure

## In collaboration with LCG




NorduGrid

Grid3/OSG

# Grid monitoring

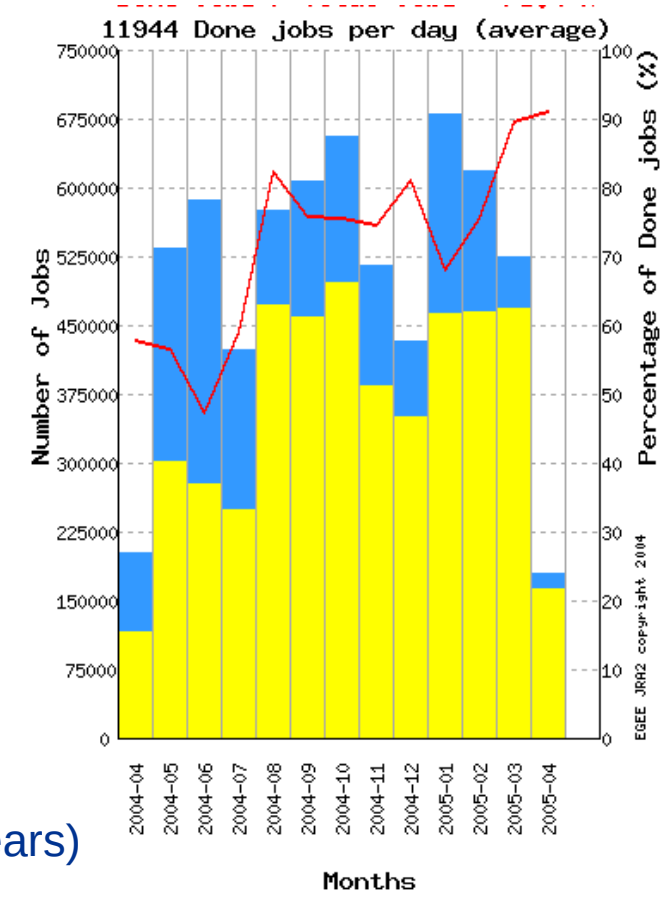
- **Operation of Production Service: real-time display of grid operations**
- **Accounting Information**
- **Selection of Monitoring tools:**
  - GIIS Monitor + Monitor Graphs
  - Sites Functional Tests
  - GOC Data Base
  - Scheduled Downtimes

  - Live Job Monitor
  - GridIce – VO + Fabric View
  - Certificate Lifetime Monitor

# Service Usage

- **VOs and users on the production service**
  - Active VOs:
    - HEP: 4 LHC, D0, CDF, Zeus, Babar
    - Biomed
    - ESR (Earth Sciences)
    - Computational chemistry
    - Magic (Astronomy)
    - EGEODE (Geo-Physics)
  - Registered users in these VO: 1000
  + Many local VOs, supported by their ROCs

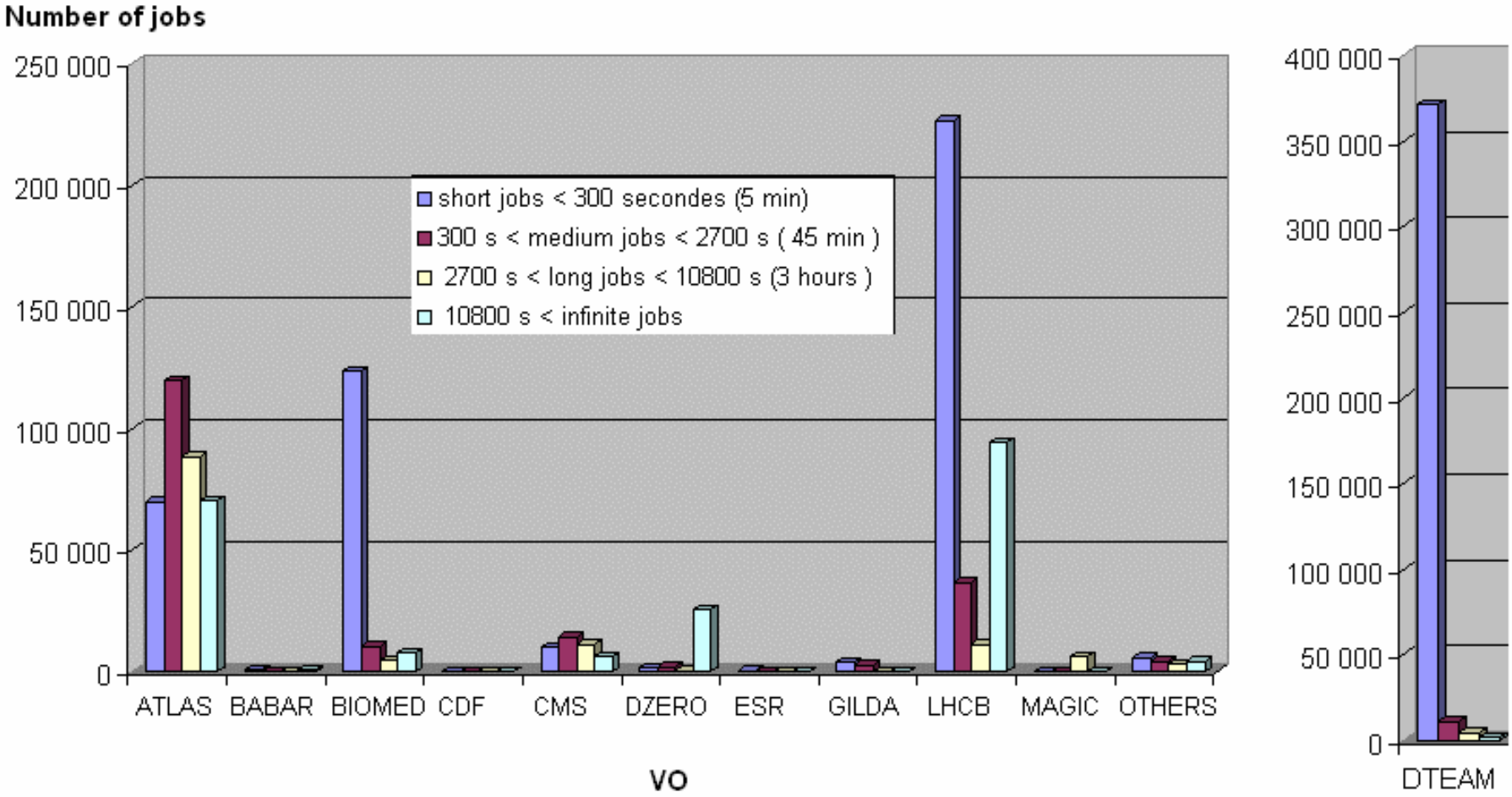
- **Scale of work performed:**
  - LHC Data challenges 2004:
    - >1 M SI2K years of CPU time (~1000 CPU years)
    - 400 TB of data generated, moved and stored
    - 1 VO achieved ~4000 simultaneous jobs (~4 times CERN grid capacity)




**Number of jobs processed per month (April 2004-April 2005)**

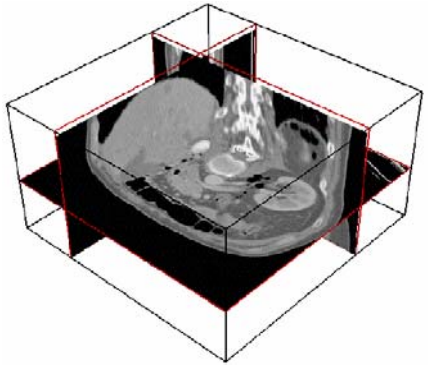
# EGEE infrastructure usage

- **Average job duration January 2005 – June 2005 for the main VOs**



# EGEE pilot applications

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

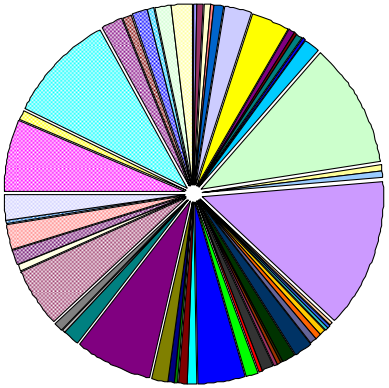


- **Biomedical Applications**
  - Similar computing and data storage requirements
  - Major additional challenge: security & privacy

# HEP success stories

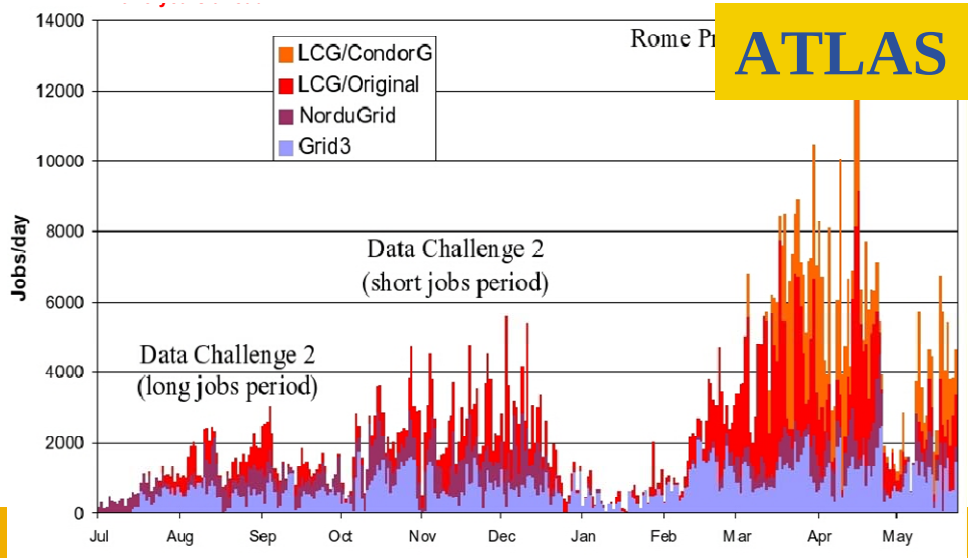
- **Fundamental activity in preparation of LHC start up**
  - Physics
  - Computing systems
- **Examples:**
  - LHCb: **~700 CPU/years in 2005** on the EGEE infrastructure
  - ATLAS: **over 10,000 jobs per day**
    - Comprehensive analysis: see S.Campana et al., "Analysis of the ATLAS Rome Production 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**



DIRAC.Barcelona.es 0.214%
DIRAC.CERN.ch 0.571%
DIRAC.CracowAgu.pl 0.001%
DIRAC.LHCBONLINE.ch 0.779%
DIRAC.PNPI.ru 0.000%
DIRAC.ScotGrid.uk 3.068%
DIRAC.Zurich.ch 0.756%
LCG.BHAM-HEP.uk 0.705%
LCG.Bari.it 1.357%
LCG.CERN.ch 10.960%
LCG.CGG.fr 0.676%
LCG.CNAF.it 13.196%
LCG.CPPM.fr 0.242%
LCG.CY01.cy 0.103%
LCG.Cambridge.uk 0.010%
LCG.Durham.uk 0.476%
LCG.FZK.de 1.708%
LCG.Firenze.it 1.047%
LCG.GR-02.gr 0.226%
LCG.GR-04.gr 0.056%
LCG.HPC2N.se 0.001%
LCG.IFCA.es 0.022%
LCG.IN2P3.fr 4.143%
LCG.IPP.bg 0.033%
LCG.Imperial.uk 0.891%
LCG.JINR.ru 0.472%
LCG.Lancashire.uk 6.796%
LCG.Manchester.uk 0.285%
LCG.Montreal.ca 0.069%
LCG.NSC.se 0.465%
LCG.Oxford.uk 1.214%
LCG.PNPI.ru 0.278%
LCG.Pisa.it 0.121%
LCG.RAL-HEP.uk 0.938%
LCG.RHUL.uk 2.168%
LCG.Sheffield.uk 0.094%
LCG.Toronto.ca 0.343%
LCG.UCL-CCC.uk 1.455%
DIRAC.Zurich-spz.ch 0.003%
LCG.ACAD.bg 0.106%
LCG.Barcelona.es 0.281%
LCG.Bologna.it 0.032%
LCG.CESGA.es 0.528%
LCG.CNAF-GRIDIT.it 0.012%
LCG.CNB.es 0.385%
LCG.CSCS.ch 0.282%
LCG.Cagliari.it 0.515%
LCG.Catania.it 0.551%
LCG.Edinburgh.uk 0.031%
LCG.Ferrara.it 0.073%
LCG.GR-01.gr 0.349%
LCG.GR-03.gr 0.171%
LCG.GRNET.gr 1.170%
LCG.ICI.ro 0.088%
LCG.IHEP.su 1.245%
LCG.INTA.es 0.076%
LCG.ITEP.ru 0.792%
LCG.Iowa.us 0.287%
LCG.KFKI.hu 1.436%
LCG.Legnaro.it 1.569%
LCG.Milano.it 0.770%
LCG.NIKHEF.nl 5.140%
LCG.Napoli.it 0.175%
LCG.PIC.es 2.366%
LCG.Padova.it 2.041%
LCG.QMUL.uk 6.407%
LCG.RAL.uk 9.518%
LCG.SARA.nl 0.675%
LCG.Torino.it 1.455%
LCG.Triumf.ca 0.105%
LCG.USC.es 1.853%

**LHCb**

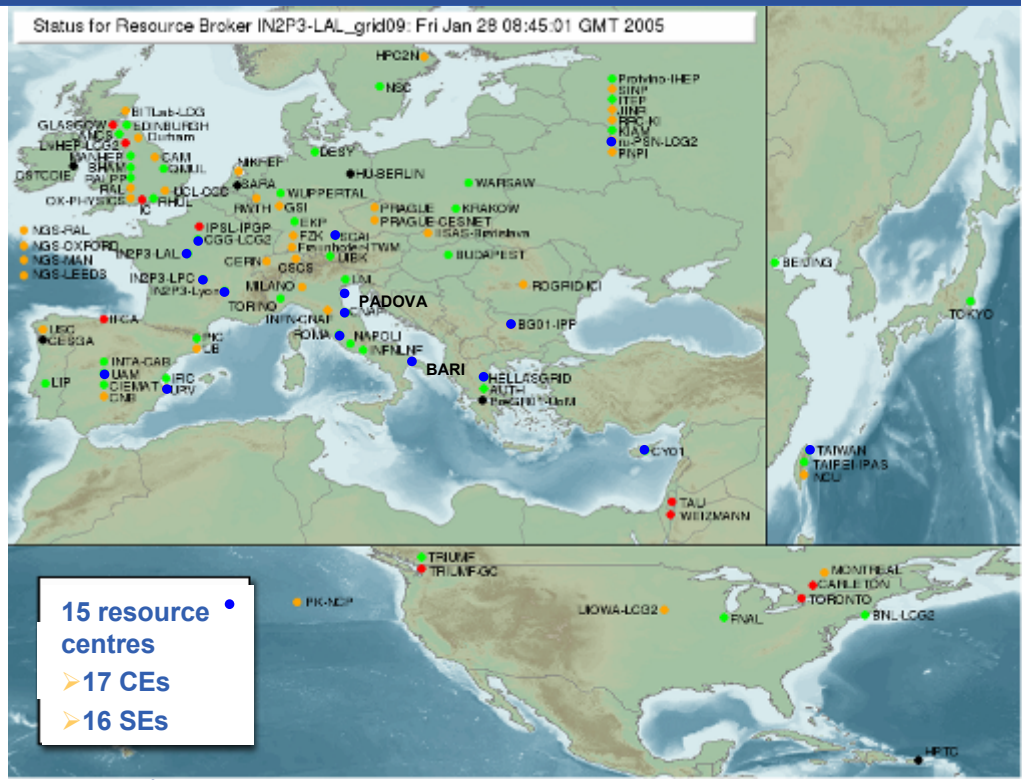


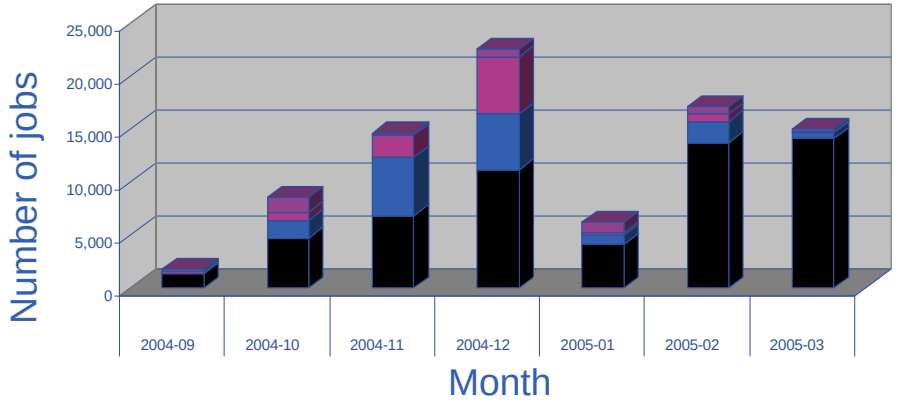

**ATLAS**

# BioMed Overview

- **Infrastructure**
  - ~3.000 CPUs
  - ~12 TB of disk
  - in 9 countries

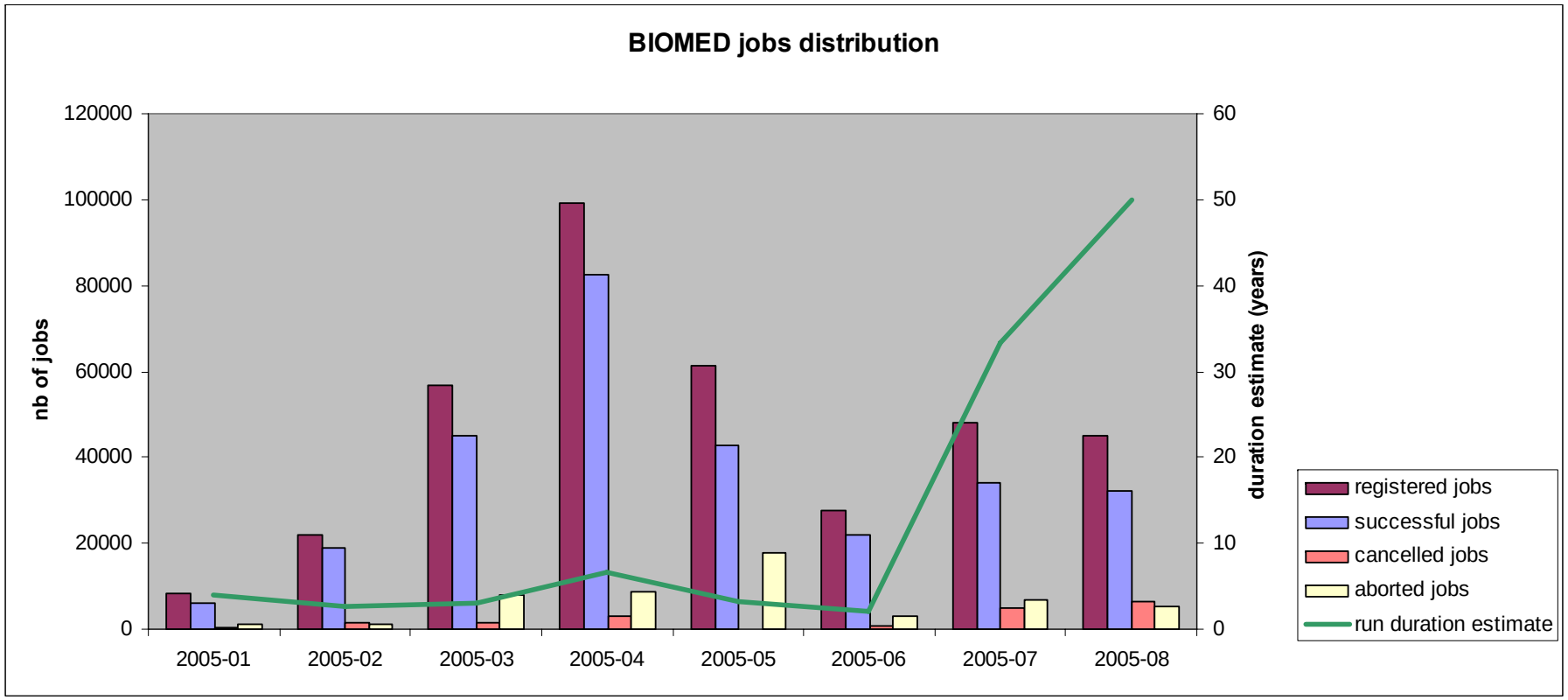
- **>50 users in 7 countries working with 12 applications**
- **18 research labs**



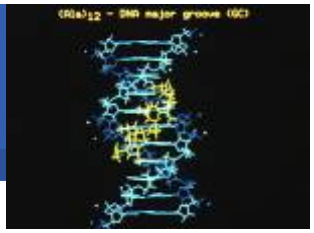



# Biomed Virtual Organisation

- ~ 70 users, 9 countries
- > 12 Applications (medical image processing, bioinformatics)
- ~3000 CPUs, ~12 TB disk space
- ~100 CPU years, ~ 500K jobs last 6 months



# Bioinformatics

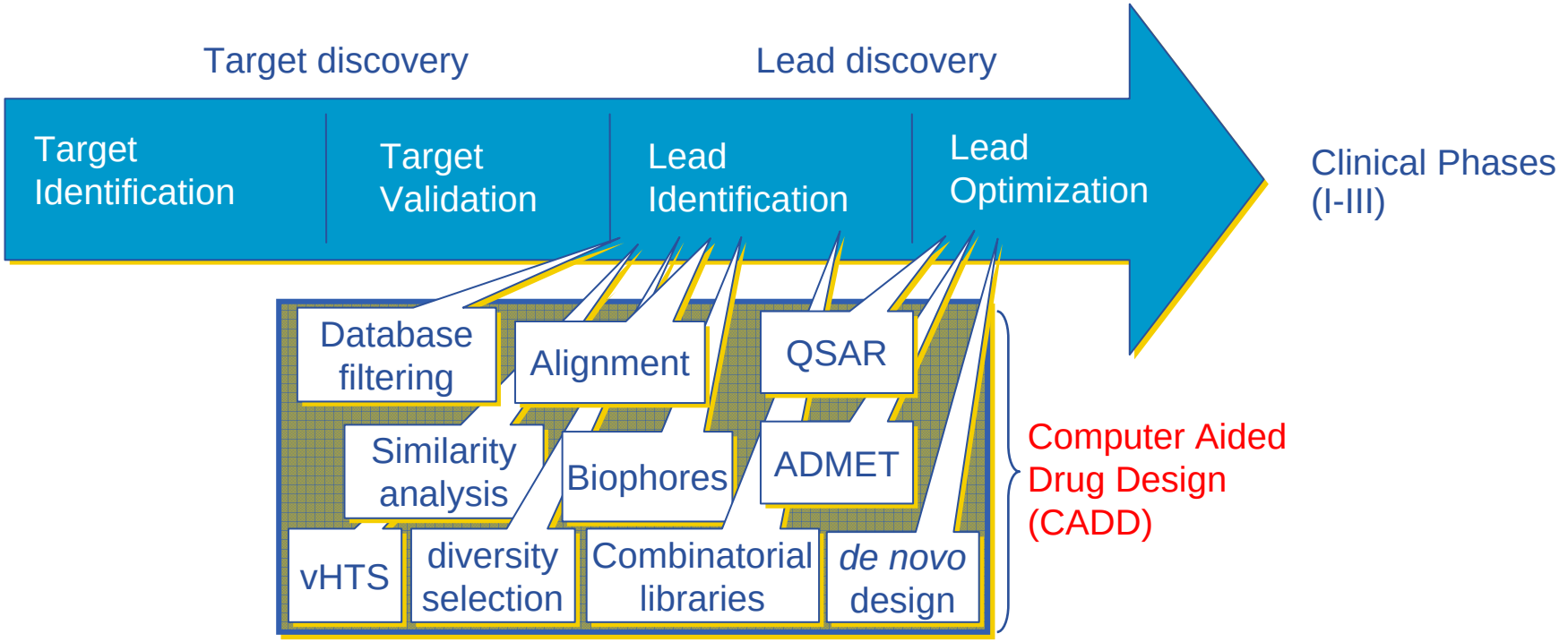


- **GPS@: Grid Protein Sequence Analysis**
  - **Gridified version of NPSA web portal**
    - Offering proteins databases and sequence analysis algorithms to the bioinformaticians (3000 hits per day)
    - Need for large databases and big number of short jobs
  - **Objective**: increased computing power
  - **Status**: 9 bioinformatic softwares gridified
  - **Grid added value**: open to a wider community with larger bioinformatic computations

- **xmipp_MLrefine**
  - **3D structure analysis of macromolecules**
    - From (very noisy) electron microscopy images
    - Maximum likelihood approach to find the optimal model
  - **Objective**: study molecule interaction and chem. properties
  - **Status**: algorithm being optimised and ported to 3D
  - **Grid added value**: parallel computation on different resources of independent jobs

# Drug Discovery

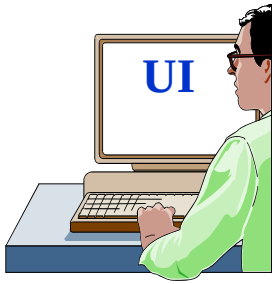
- **Demonstrate the relevance and the impact of the grid approach to address Drug Discovery for neglected diseases**



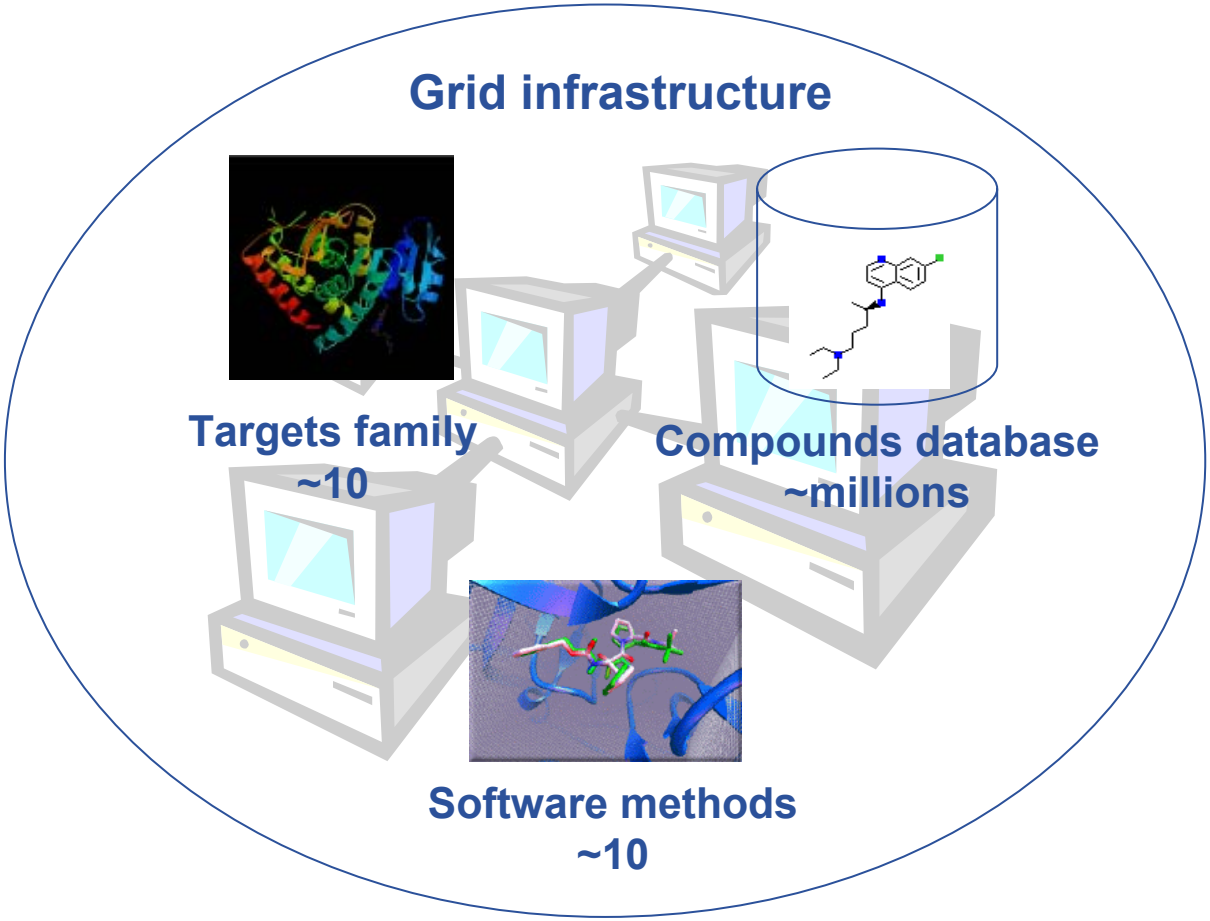

Duration: 12 – 15 years, Costs: 500 - 800 million US $

# Docking platform components

- **Predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure**

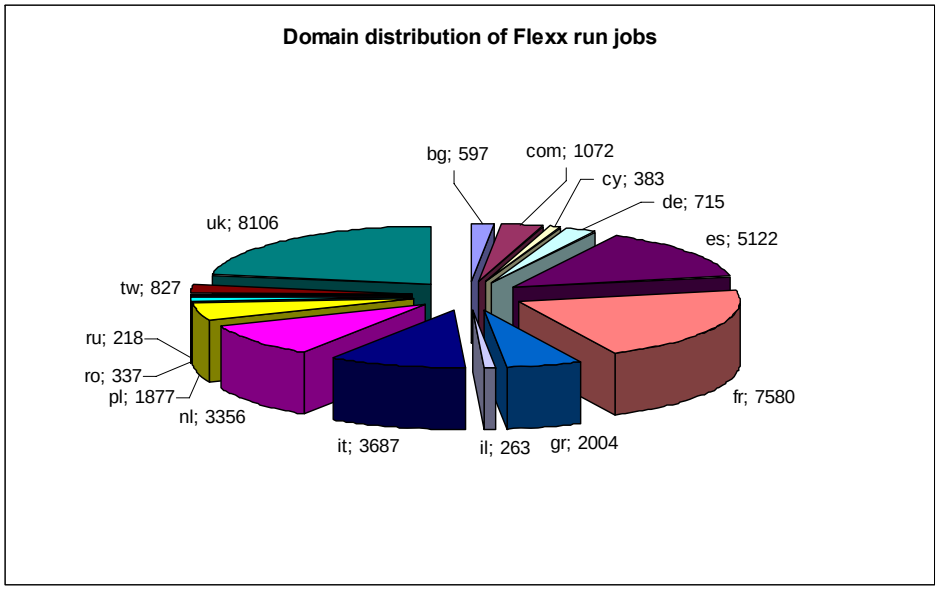



**Parameter / scoring settings**



# First biomedical data challenge: World-wide In Silico Docking On Malaria (WISDOM)

- **Significant biological parameters**
  - two different molecular docking applications (Autodock and FlexX)
  - about one million virtual ligands selected
  - target proteins from the parasite responsible for malaria
- **Significant numbers**
  - Total of about 46 million ligands docked in 6 weeks
  - 1TB of data produced
  - Up 1000 computers in 15 countries used simultaneously corresponding to about 80 CPU years




**WISDOM open day**

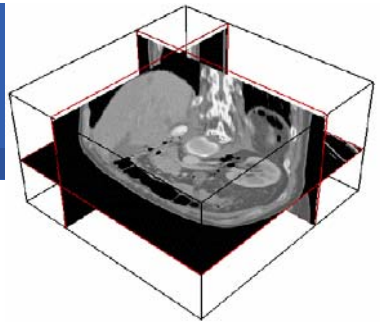
**December 16th, 2005, Bonn (Germany)**

**Discuss Data Challenge results**

**Prepare next steps towards a malaria Grid (EGEE-II, Embrace, Bioinfogrid)**

**Information: http://wisdom.eu-egee.fr**

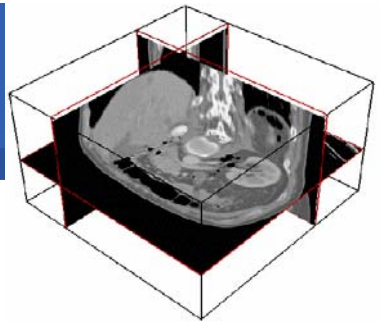
# Medical imaging



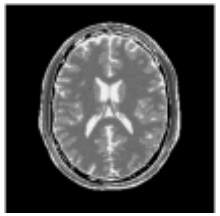
- **GATE**
  - **Radiotherapy planning**
    - Improvement of precision by Monte Carlo simulation
    - Processing of DICOM medical images
  - **Objective**: very short computation time compatible with clinical practice
  - **Status**: development and performance testing
  - **Grid Added Value**: parallelisation reduces computing time

- **CDSS**
  - **Clinical Decision Support System**
    - Assembling knowledge databases
    - Using image classification engines
  - **Objective:** access to knowledge databases from hospitals
  - **Status**: from development to deployment, some medical end users
  - **Grid Added Value**: ubiquitous, managed access to distributed databases and engines

# Medical imaging

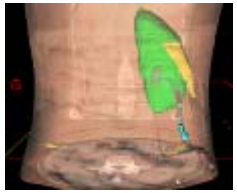


- **SiMRI3D**
  - **3D Magnetic Resonance Image Simulator**
    - MRI physics simulation, parallel implementation
    - Very compute intensive



  - **Objective**: offering an image simulator service to the research community
  - **Status**: parallelised and now running on EGEE resources
  - **Grid Added Value**: enables simulation of high-res images

- **gPTM3D**
  - **Interactive tool to segment and analyse medical images**
    - A non gridified version is distributed in several hospitals
    - Need for very fast scheduling of interactive tasks



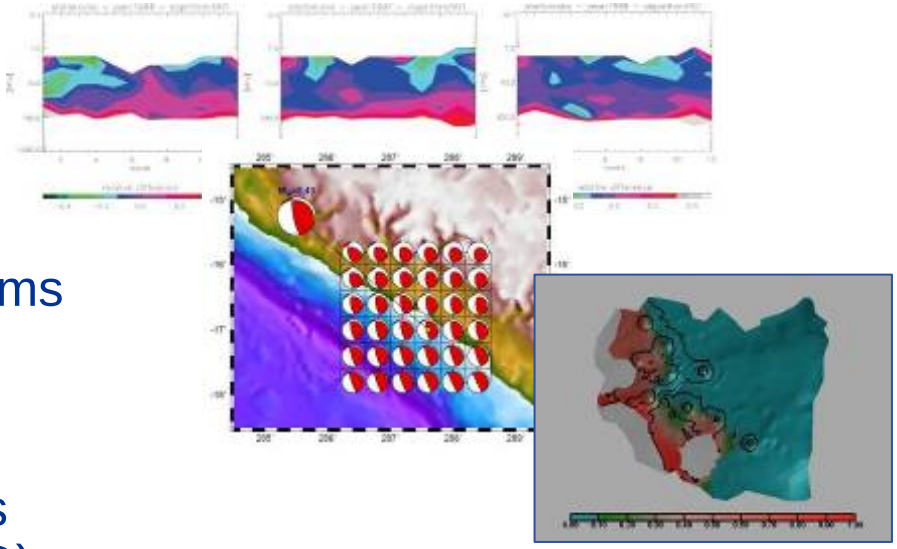
  - **Objectives**: shorten computation time using the grid
    - Interactive reconstruction time: < 2min and scalable
  - **Status**: development of the gridified version being finalized
  - **Grid Added Value**: permanent availability of resources

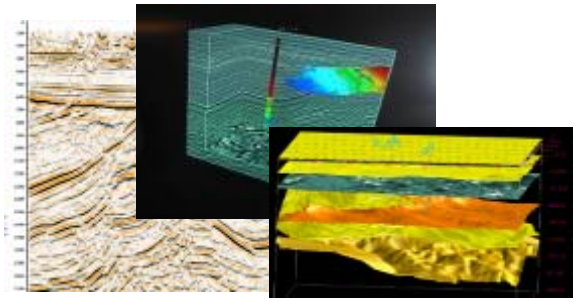
# Generic Applications

- **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 (earth observation, geophysics, hydrology, seismology)
    - MAGIC (astrophysics)
    - Computational Chemistry
    - PLANCK (astrophysics and cosmology)
    - Drug Discovery
    - E-GRID (e-finance and e-business)
    - GRACE (grid search engine, ended 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

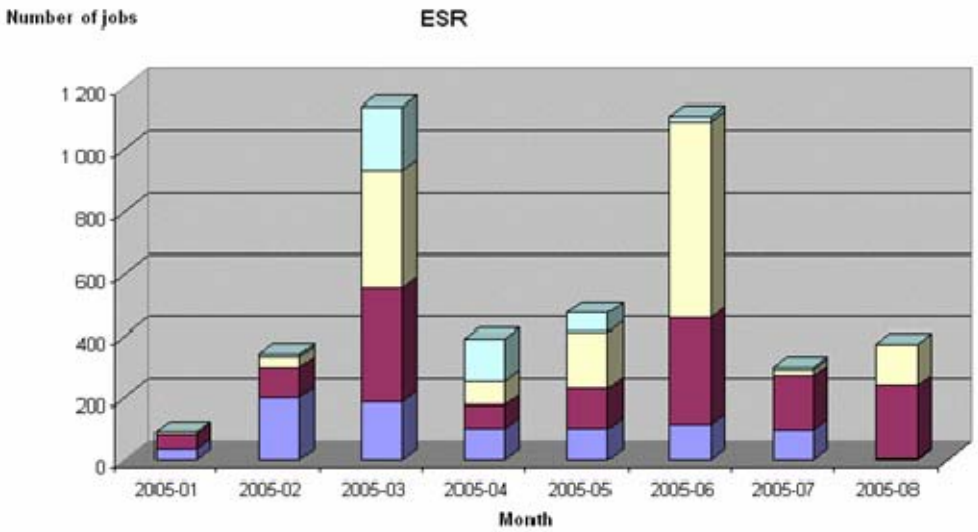


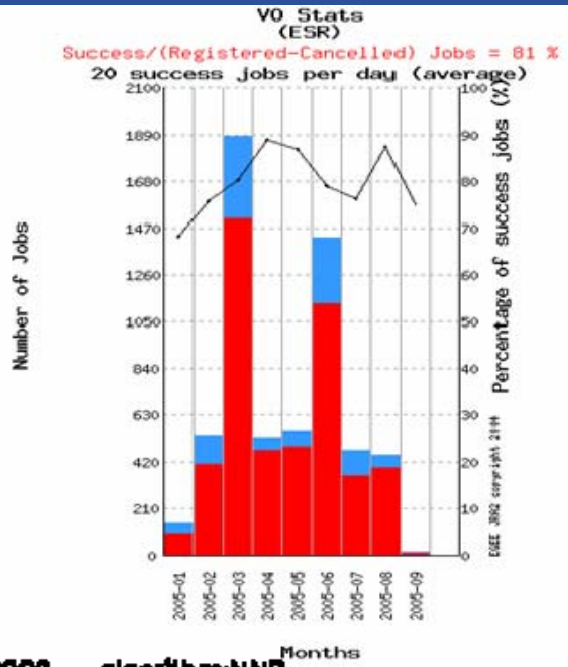


- **A large variety of applications ported on EGEE which incites new users**
- **Interactive Collaboration of the teams around a project**

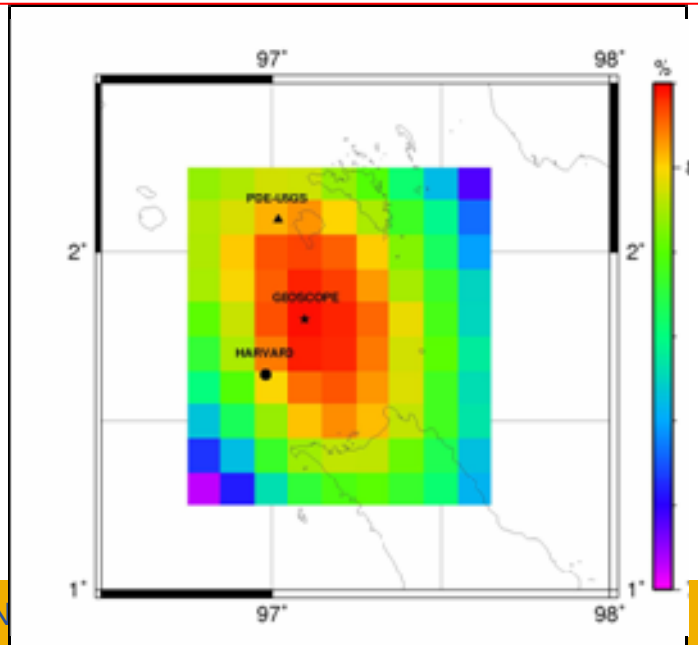
# Generic Applications' use of EGEE

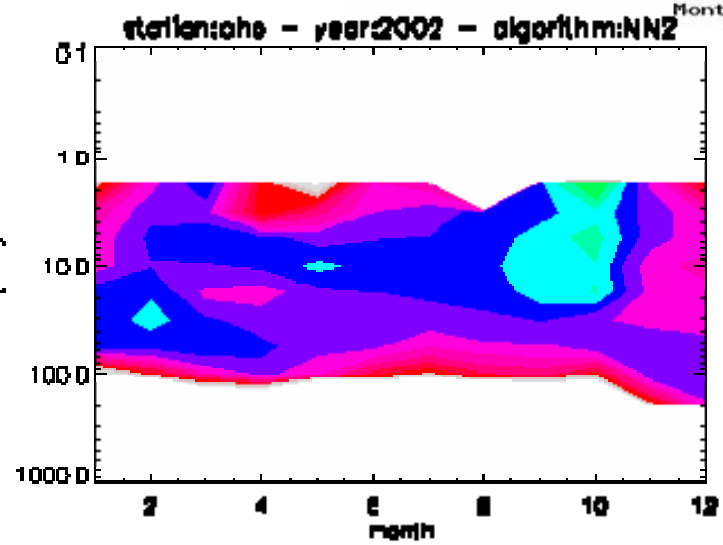
Enabling Grids for E-sciencE




Earthquakes' epicenter determination



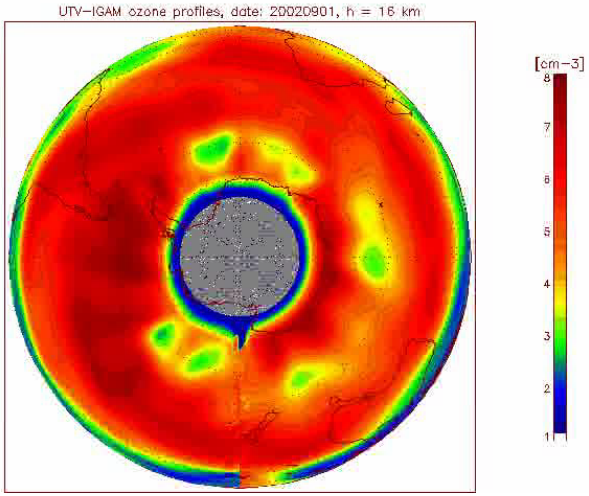



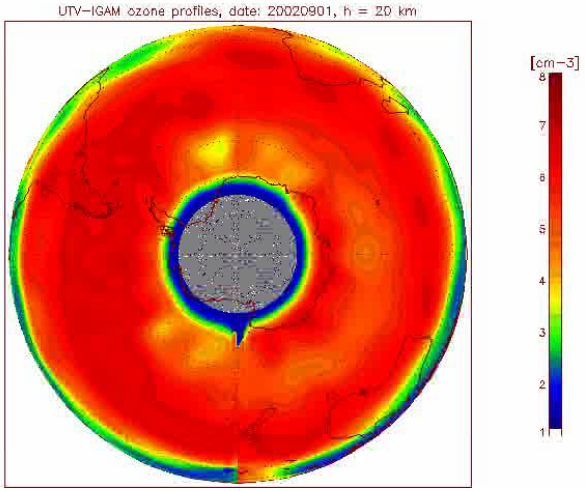

Ozone maps

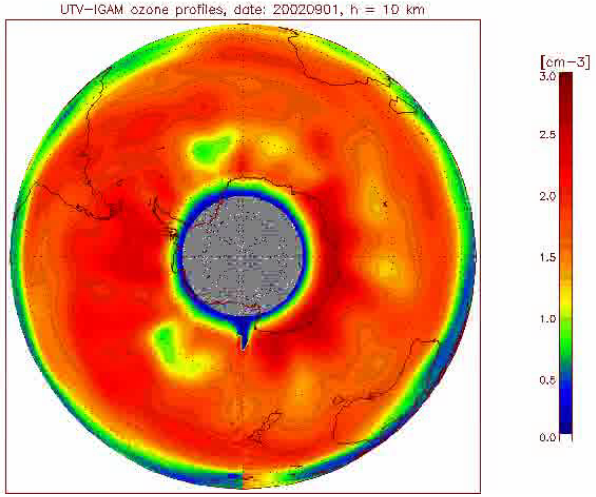
Climate

# Ozone profiles from GOME (1997-2003)

S. Casadio – ESA ESRIN
(GOME 3D Ozone volume over Antarctica - Sept 02, NNO Level 2 products generate in EDG)

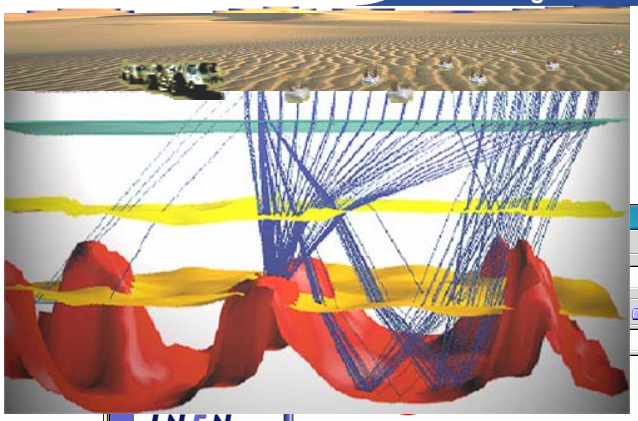


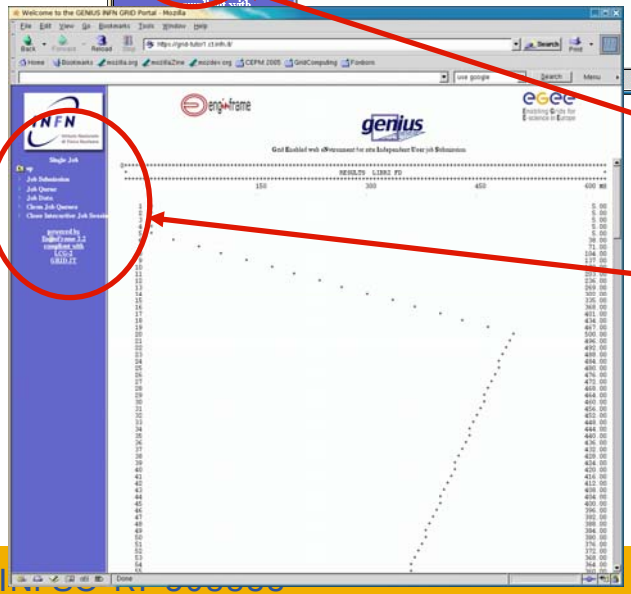



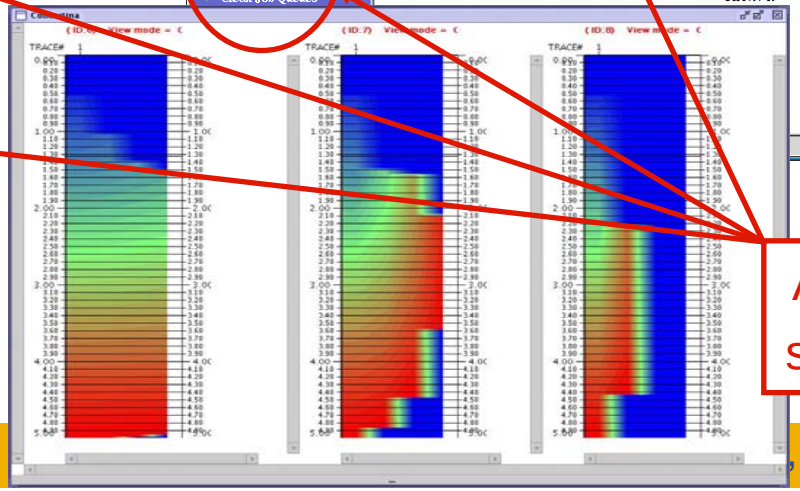



# Earth Science (industrial): EGEODE example

Enabling Grids for E-sciencE



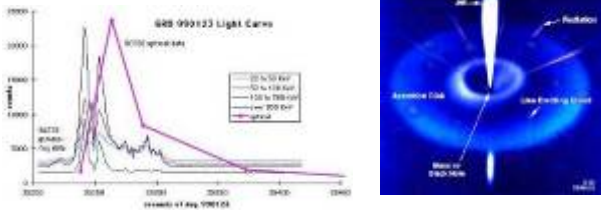


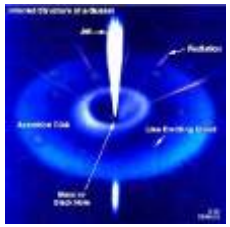


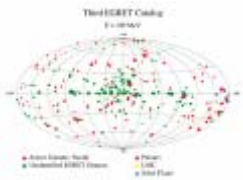
Application-specific services in GENIUS

# MAGIC

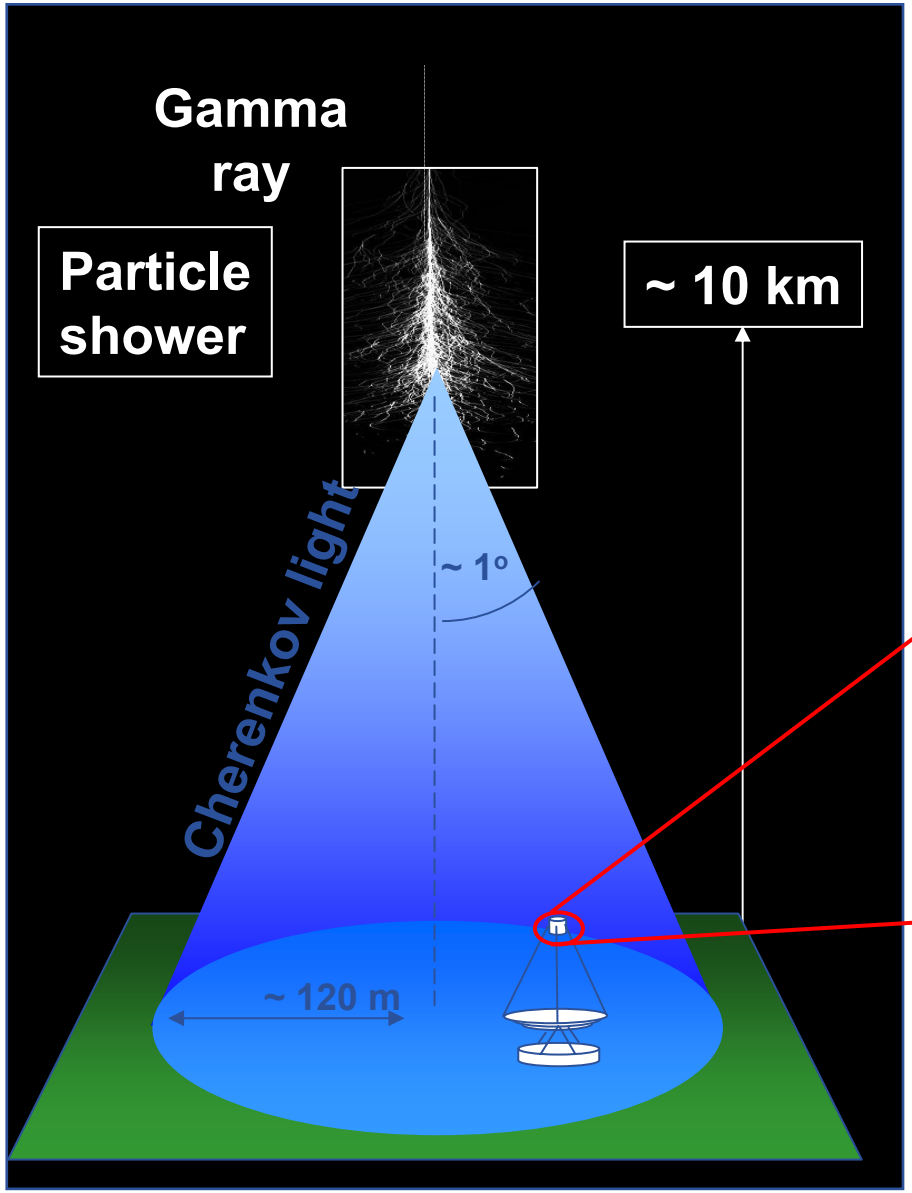
- **Ground based Air Cerenkov Telescope 17 m diameter**
- **Physics Goals:**
  - Origin of VHE Gamma rays
  - Active Galactic Nuclei
  - Supernova Remnants
  - Unidentified EGRET sources
  - Gamma Ray Burst
- **MAGIC II will come 2007**
- **Grid added value**
  - Enable “(e-)scientific” collaboration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



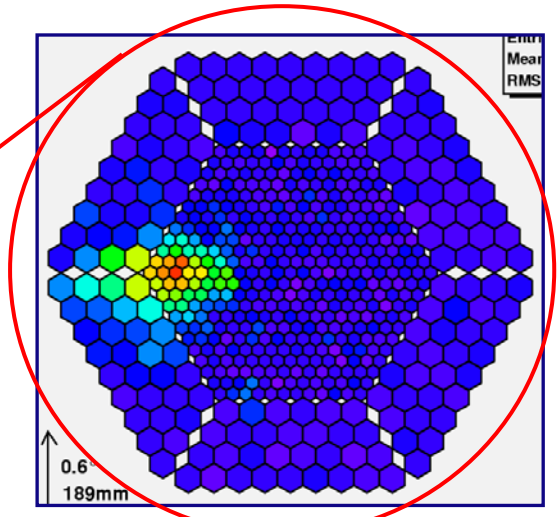




# Ground based γ-ray astronomy



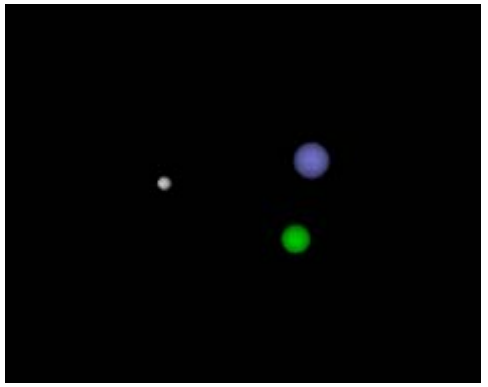

Cherenkov light Image of particle shower in telescope camera

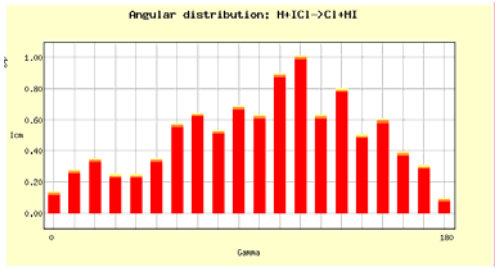


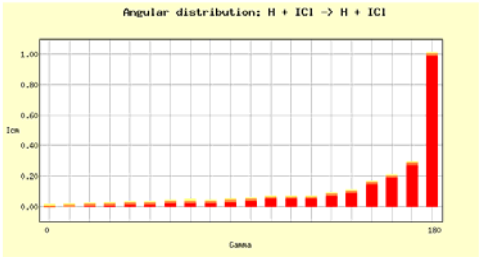

reconstruct:
arrival direction, energy
reject hadron background

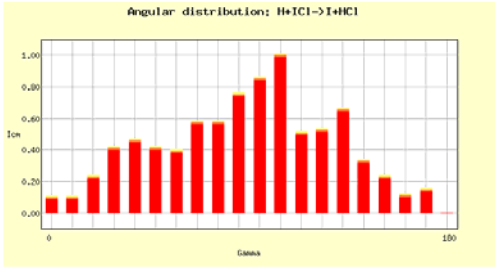
# Computational Chemistry

- **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



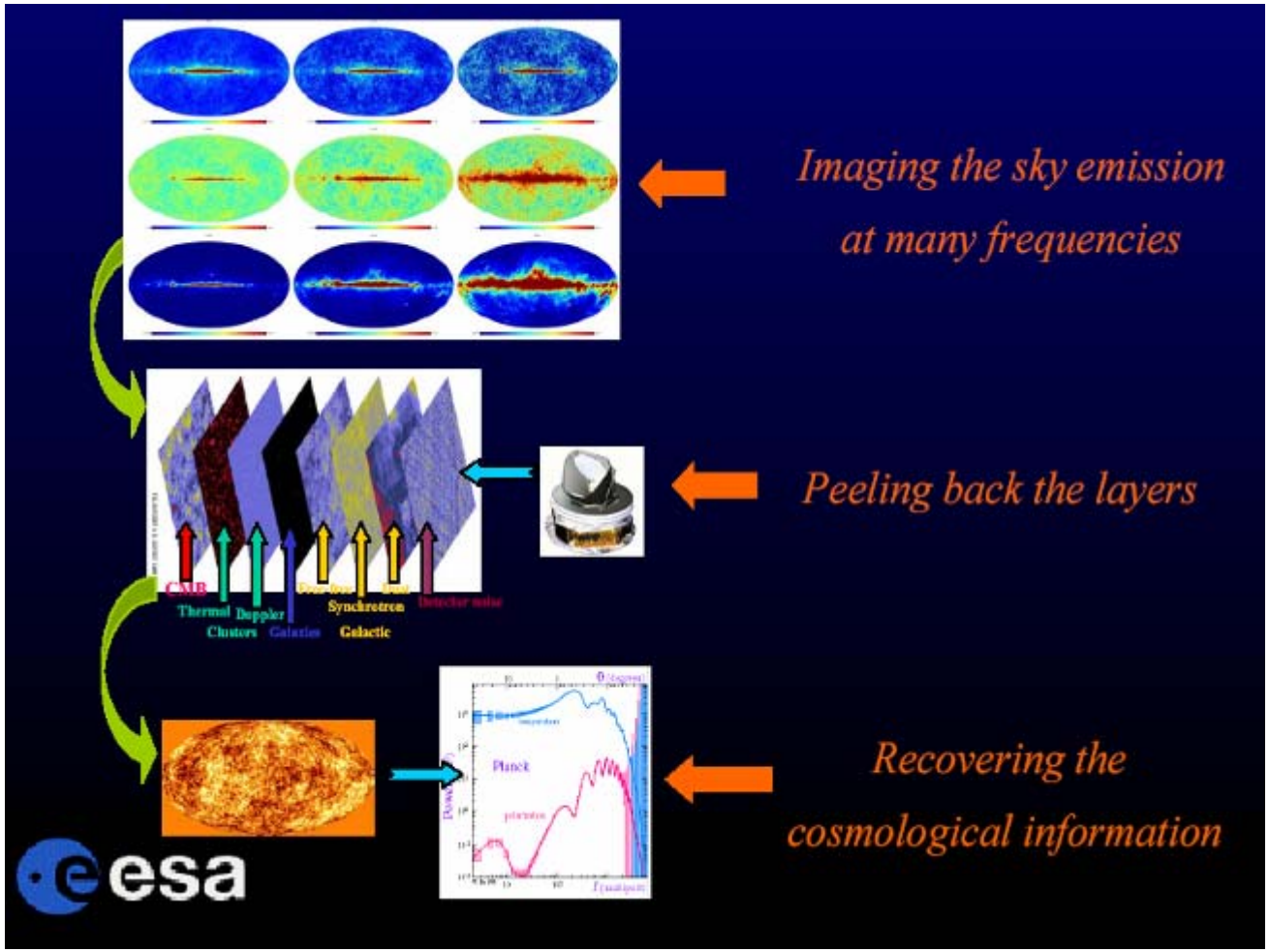


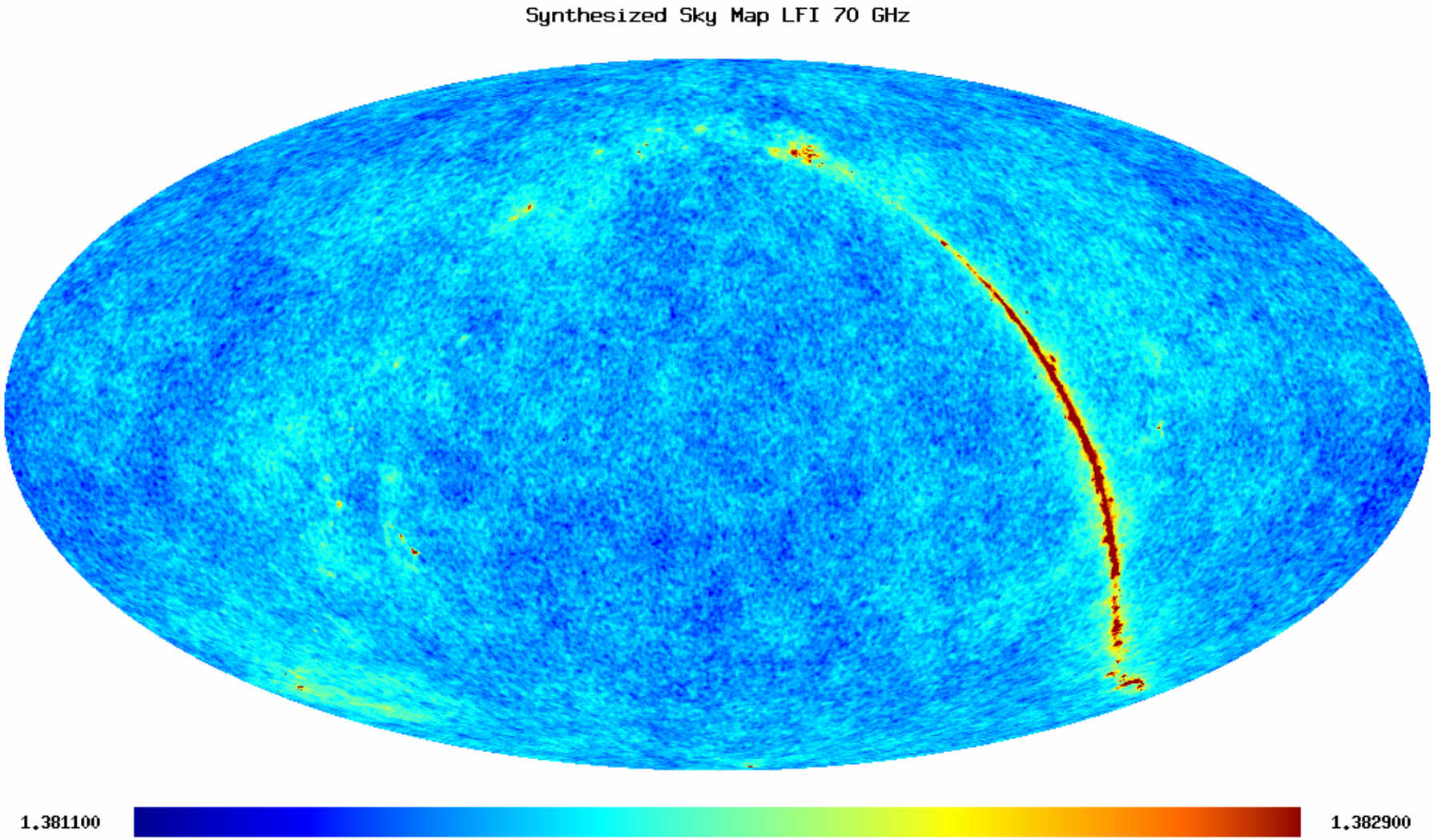



# Planck

- **On the Grid:**
  > 12 time faster
  (only ~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



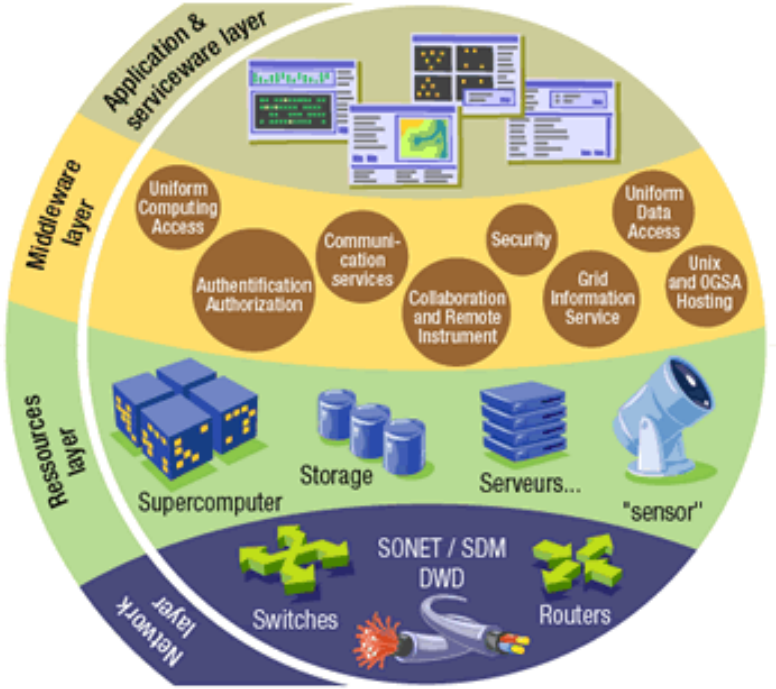
# Planck first tests



# Grid 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

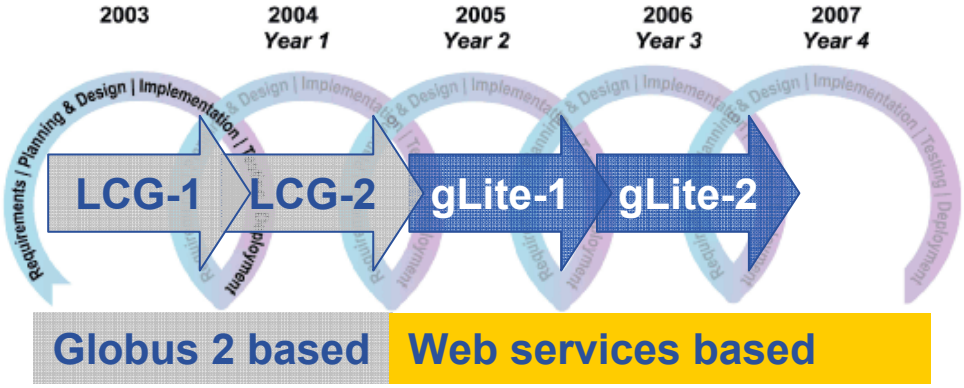


# EGEE Middleware gLite

- First release of gLite end of March 2005
  - Focus on providing users early access to prototype
  - Release 1.1 in May 05
  - Release 1.2 in July 05
  - Release 1.3 in August 05
  - Release 1.4 in October
  - see www.gLite.org
- Interoperability & Co-existence with deployed infrastructure
- Robust: Performance & Fault Tolerance
- Service oriented approach
- Open source license

# EGEE Middleware

- 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/

# Architecture & Design

- **Design team includes**
  - Representatives from middleware providers (AliEn, Condor, EDG, Globus,…)
  - Colleagues from the Operations activity
  - Partners from related projects (e.g. OSG)

- **gLite development takes into account input and experiences from applications, operations, related projects**
  - Effective exchange of ideas, requirements, solutions and technologies
  - Coordinated development of new capabilities
  - Open communication channels
  - Joint deployment and testing of middleware
  - Early detection of differences and disagreements

gLite is not "just" a software stack, it is a "new" framework for international collaborative middleware development

# User information & support

- **More than 170 training events across many countries**
  - >2500 people trained
    induction; application developer; advanced; retreats
  - Material archive online with >2000 presentations

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

- **4 conferences organized**
  - ~ 300 @ Cork
  - ~ 400 @ Den Haag
  - ~ 450 @ Athens
  - ~ 460 @ Pisa

# 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

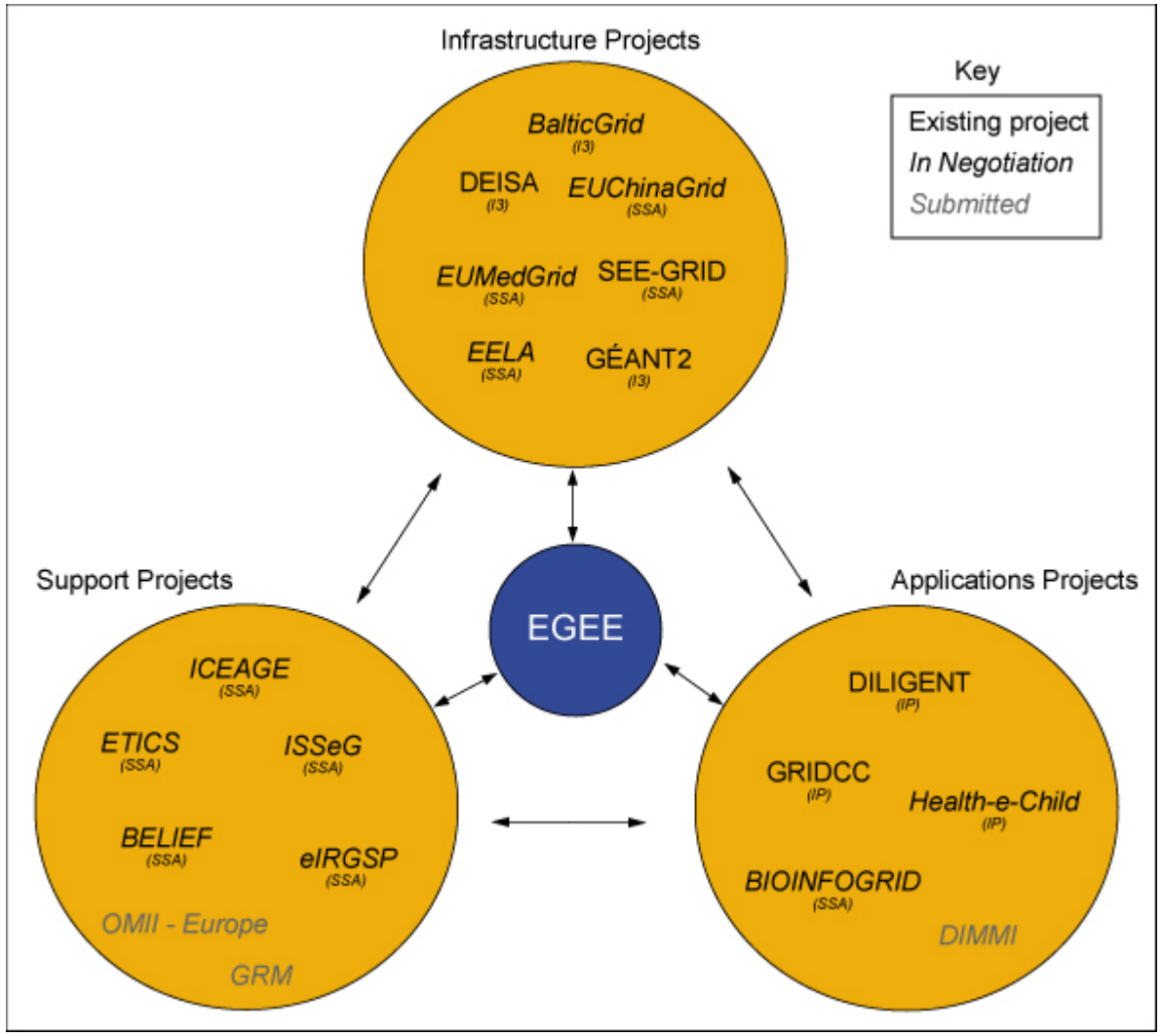
- Ongoing collaborations
  - with non-EU partners: US, Israel, Russia, Korea, Taiwan…
    - MoU with the Chonnam–Kangnung–Sejong–Collaboration project (CKSC)
    - Strong relationship KISTI (Korea Institute of Science and Technology Information), developing into partnership for EGEE II
  - with other European projects, in particular:
    - GÉANT
    - DEISA
    - SEE-GRID
  - with non-European projects:
    - OSG: OpenScienceGrid (USA)
    - NAREGI (Japan)
    - International Grid Trust Federation
      - *EU-GridPMA joining with Asia-Pacific and American counterparts*

- EGEE as incubator
  - 18 recently submitted EU proposals supported
  - More proposals in next calls and national funding programmes

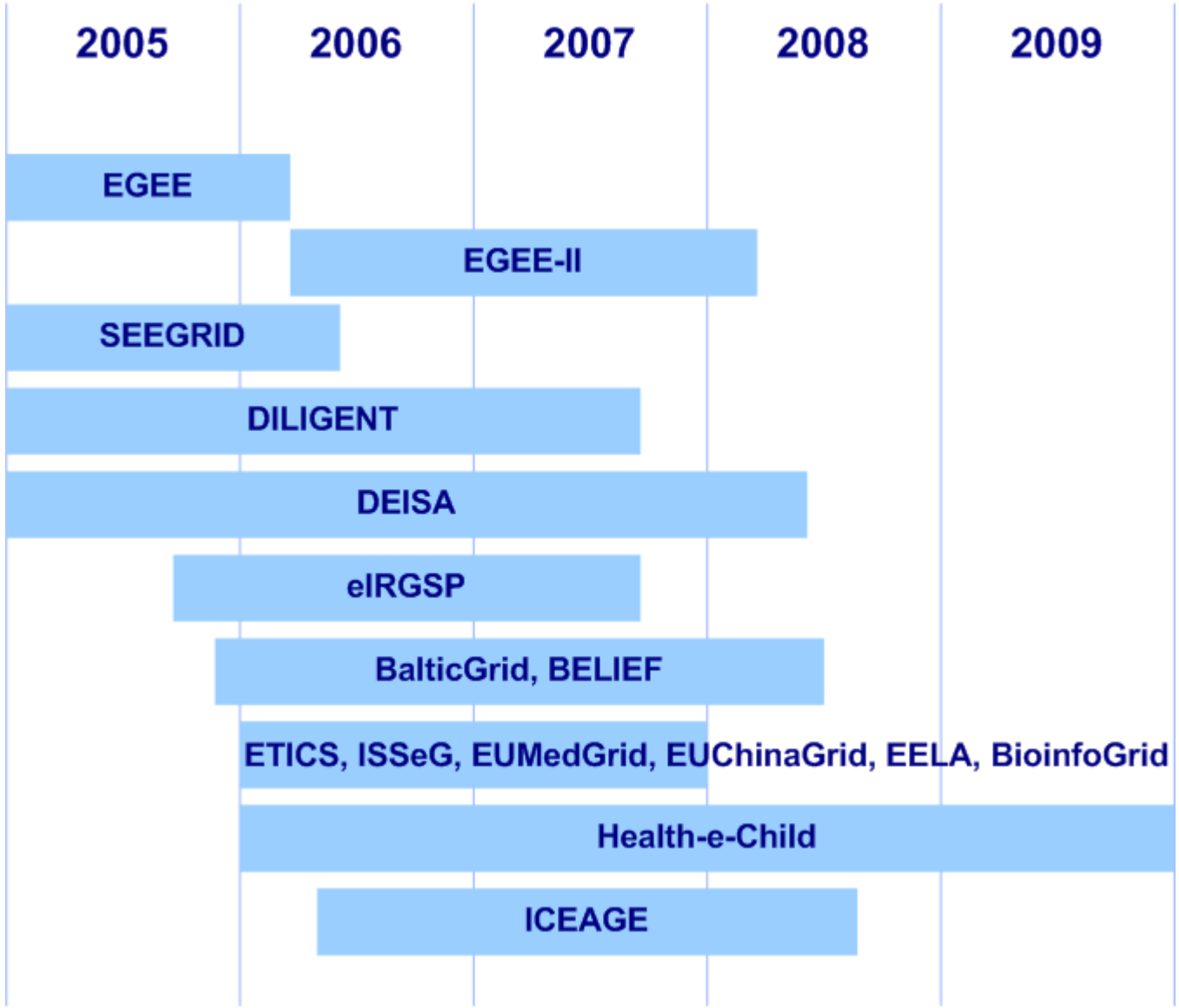
# Related projects under negotiation

| 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 |
| **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 |

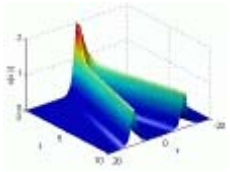
# Related Projects



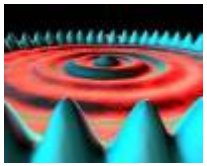
# Time scales


2005
2006
2007
2008
2009
EGEE
EGEE-II
SEEGRID
DILIGENT
DEISA
eIRGSP
BalticGrid, BELIEF
ETICS, ISSeG, EUMedGrid, EUChinaGrid, EELA, BioinfoGrid
Health-e-Child
ICEAGE

# From Phase I to II

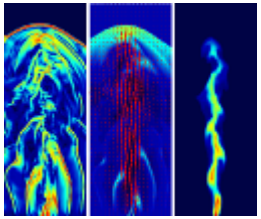


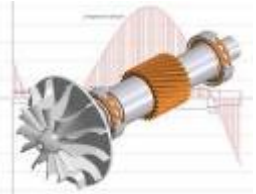
- **From 1st EGEE EU Review in February 2005:**
  - "The reviewers found the overall performance of the project very good."
  - "… remarkable achievement to set up this consortium, to realize appropriate structures to provide the necessary leadership, and to cope with changing requirements."

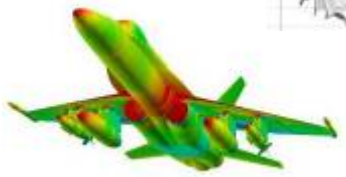


- **EGEE I**
  - Large scale deployment of EGEE infrastructure to deliver production level Grid services with selected number of applications

- **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**
      (Asia-Pacific is already a partner!)







# Conclusions

- Grids are a powerful new tool for science – as well as other fields
- Grid computing has been chosen by CERN and HEP as the most cost effective computing model
- Several other applications are already benefiting from Grid technologies (biomedical is a good example)
- Investments in grid projects are growing world-wide
- Europe is strong in the development of Grids also thanks to the success of EGEE and related projects

# Contacts

- **EGEE Website**
  http://www.eu-egee.org
- **How to join**
  http://public.eu-egee.org/join/
- **How to test**
  https://gilda.ct.infn.it
- **EGEE Project Office**
  project-eu-egee-po@cern.ch