Introduction to EGEE

Fabrizio Gagliardi
Project Director EGEE
CERN, Switzerland

EGEE tutorial, Seoul, 29 August 2005

www.eu-egee.org

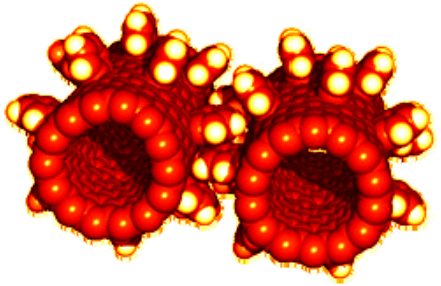
Information Society

Presentation overview

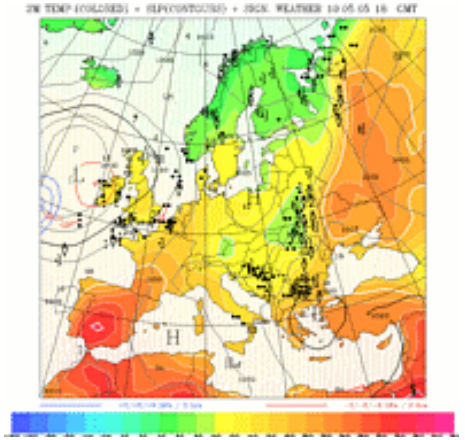
- Data intensive science and rationale for Grid computing
- Particle physics and bio-informatics examples
- General description of the EGEE project and relations to HEP CERN LCG project
- EGEE operates a production infrastructure:
  - Operations
  - Middleware
  - Applications
- Establish new user communities
- Promote and enable international collaboration

Computing intensive science

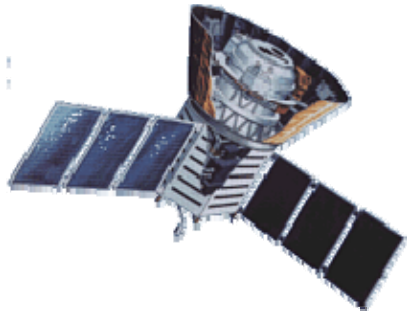
- **Science is becoming increasingly digital and needs to deal with increasing amounts of data**



- **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

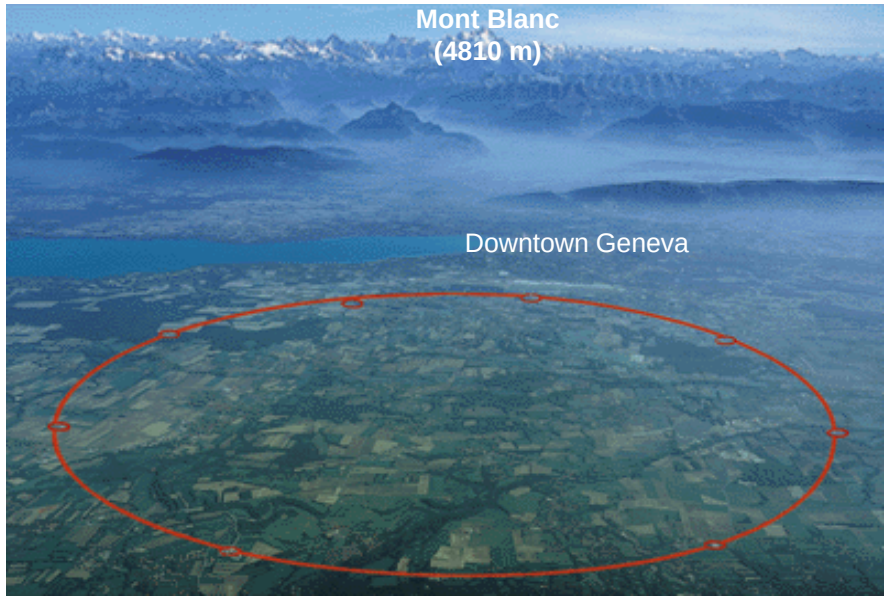


- **Experimental Science uses ever more sophisticated sensors to make precise measurements**
  - → Need high statistics
  - → Huge amounts of data
  - → Serves user communities around the world



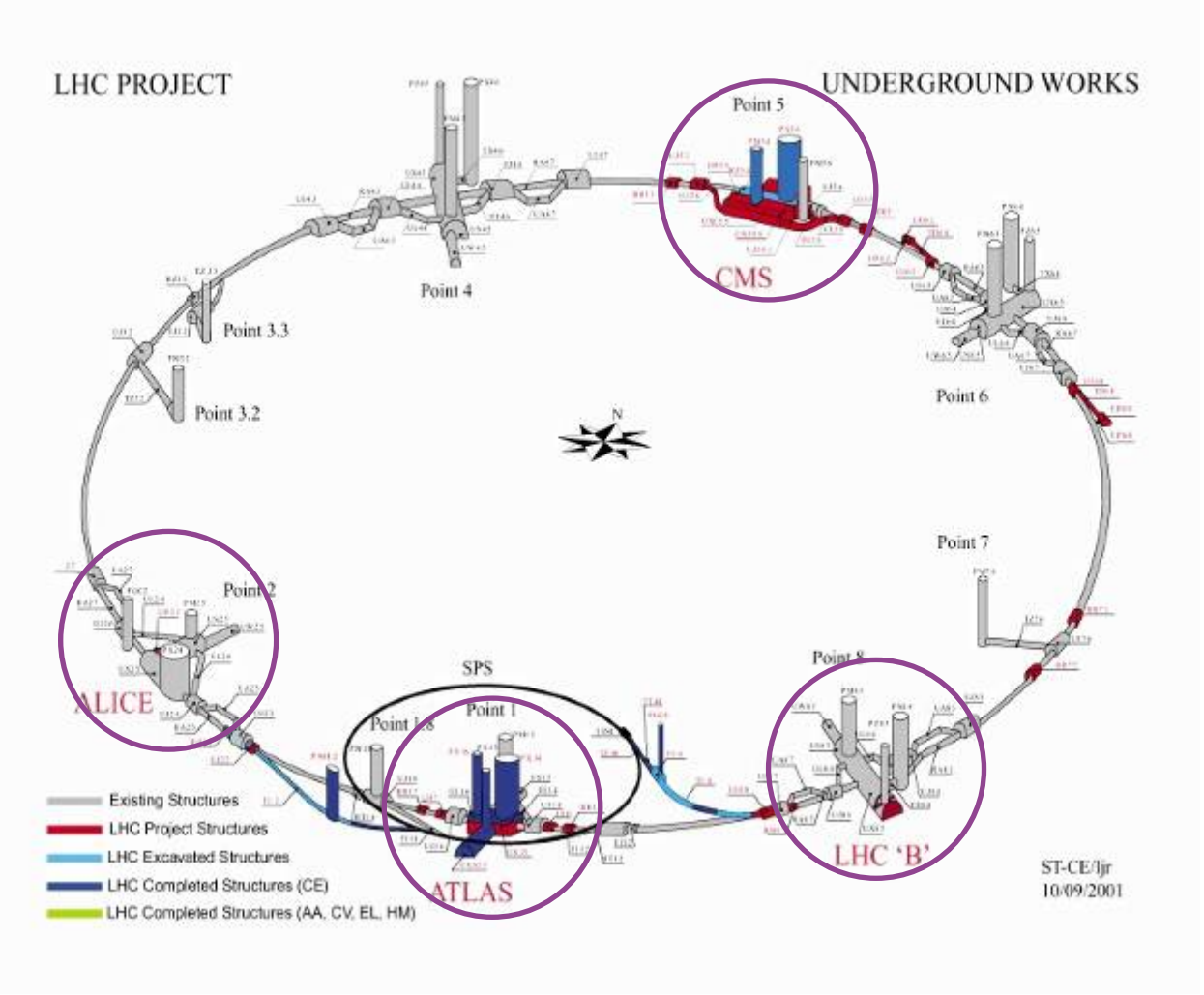
A good example: Particle Physics

- **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



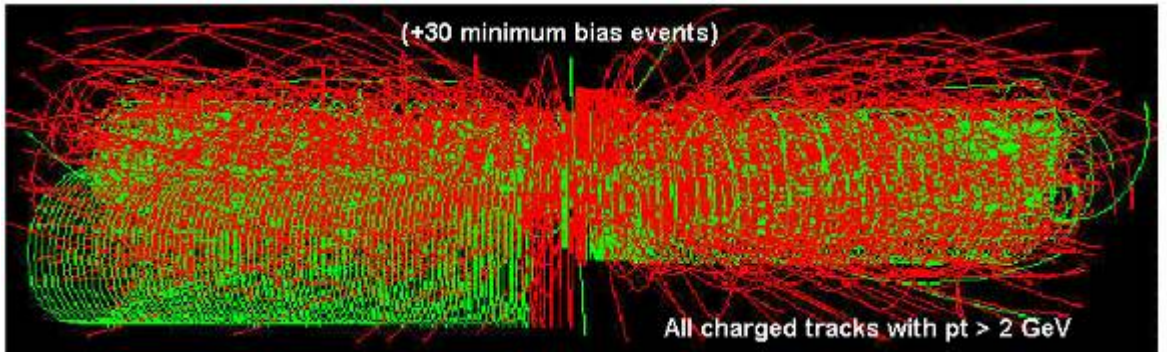
The LHC Experiments

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

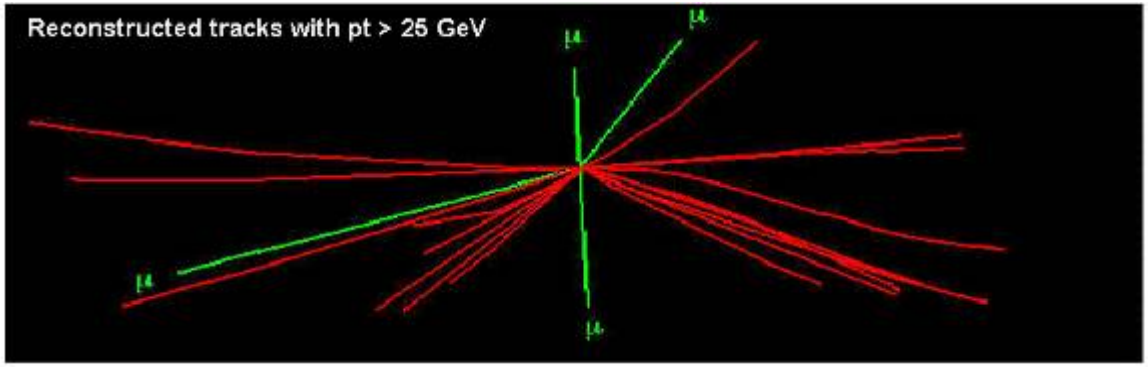


The LHC Data Challenge

Starting from this event



Looking for this "signature"

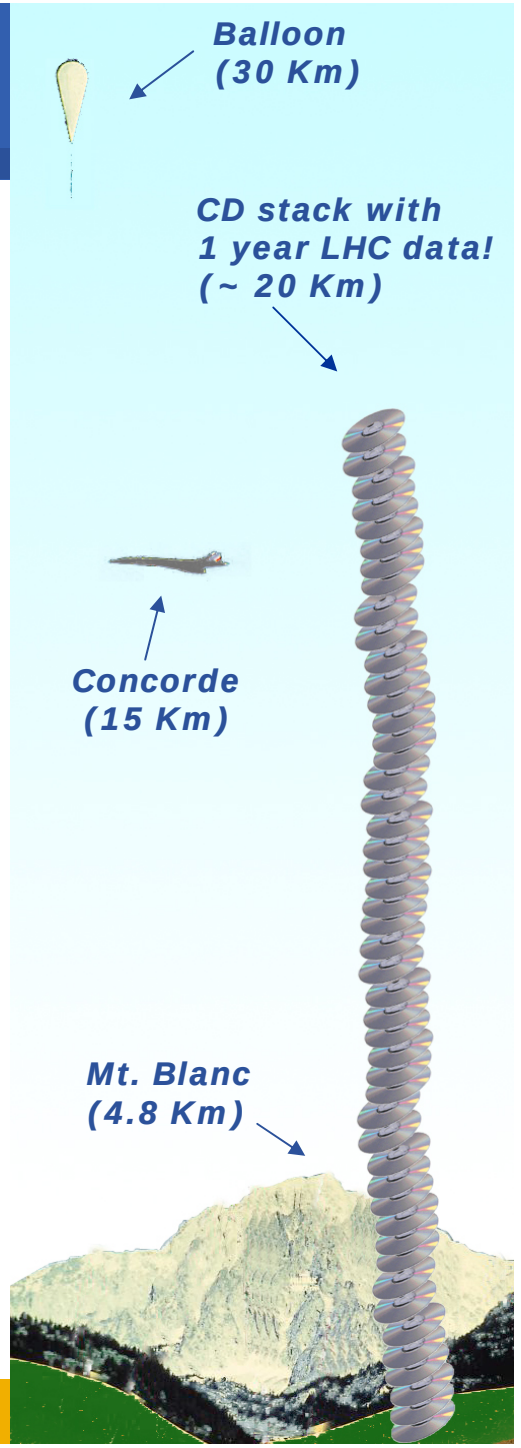


→ **Selectivity: 1 in 10^{13}**

(Like looking for a needle in 20 million haystacks)

LHC Data

- 40 million collisions per second
- After filtering, 100 collisions of interest per second
- A Megabyte of data for each collision = recording rate of 0.1 Gigabytes/sec
- 10^{10} collisions recorded each year
- ~ 10 Petabytes/year of data
- LHC data correspond to about 20 million CDs each year!
- ~ 100,000 of today's fastest PC processors



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

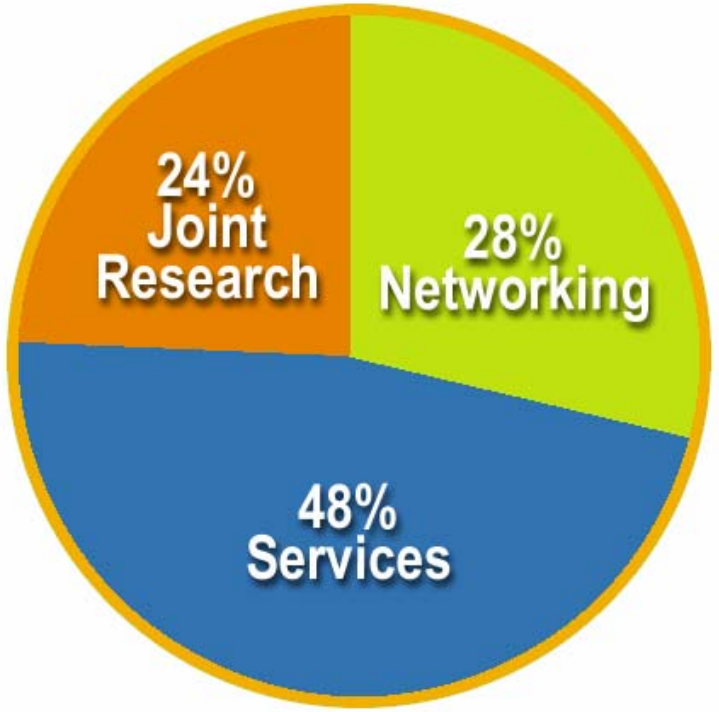
The largest e-Infrastructure: EGEE

- **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

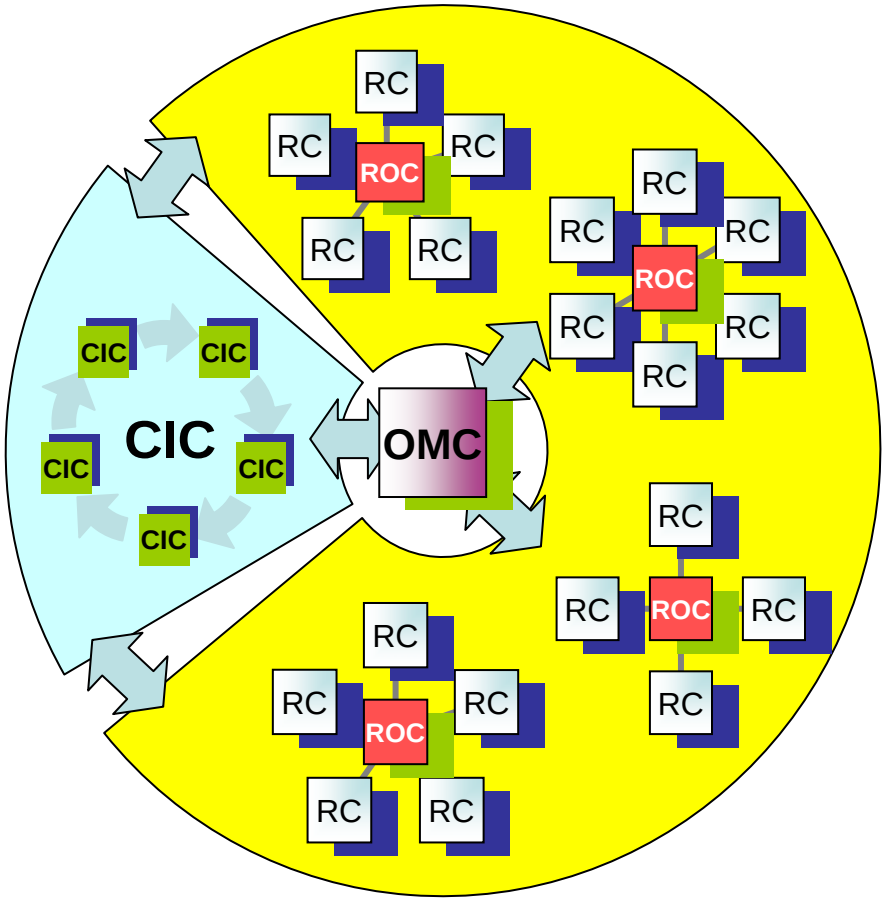
EGEE Activities

- 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

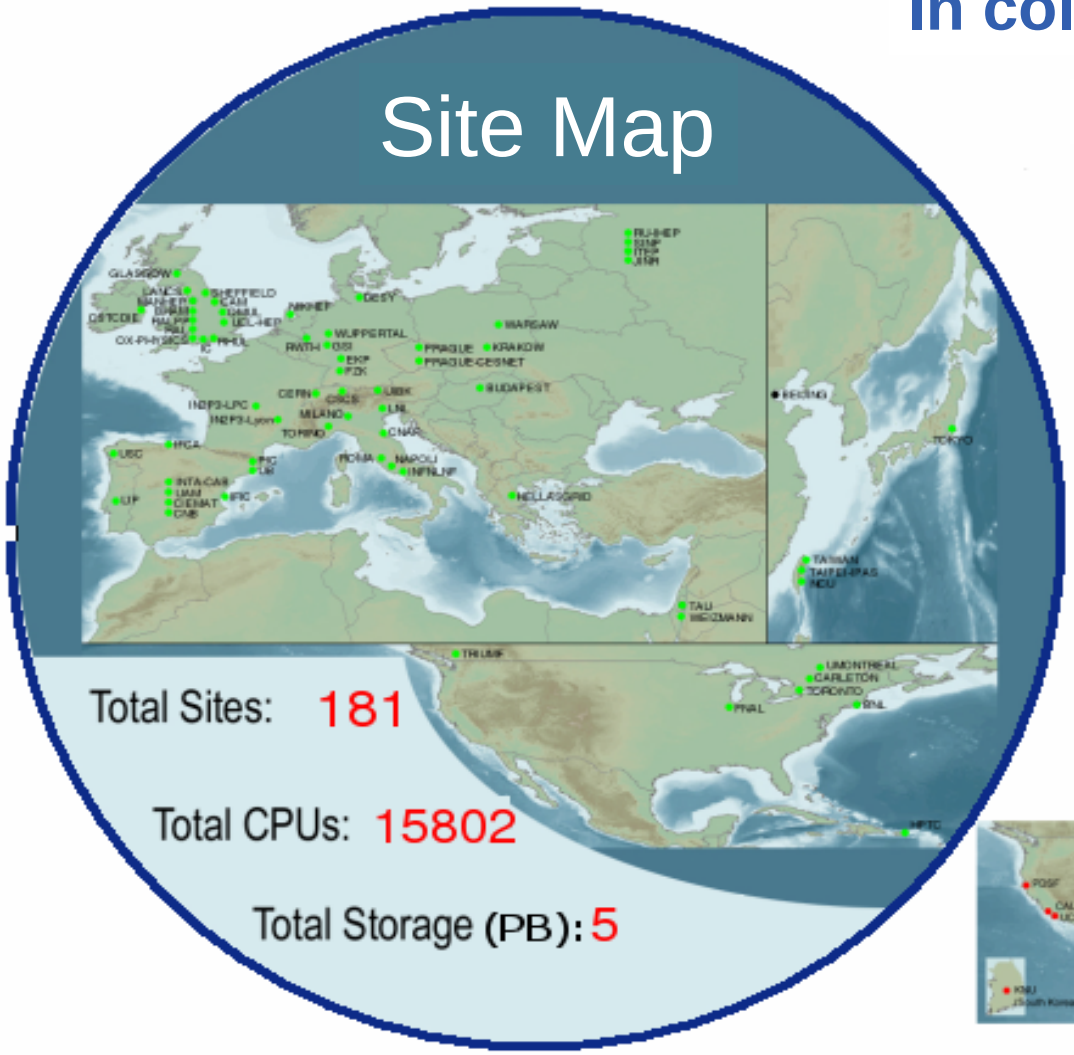


RC = Resource Centre
ROC = Regional Operations Centre
CIC = Core Infrastructure Centre
OMC = Operations Management Centre

- **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

Status 25 July 2005

Grid monitoring

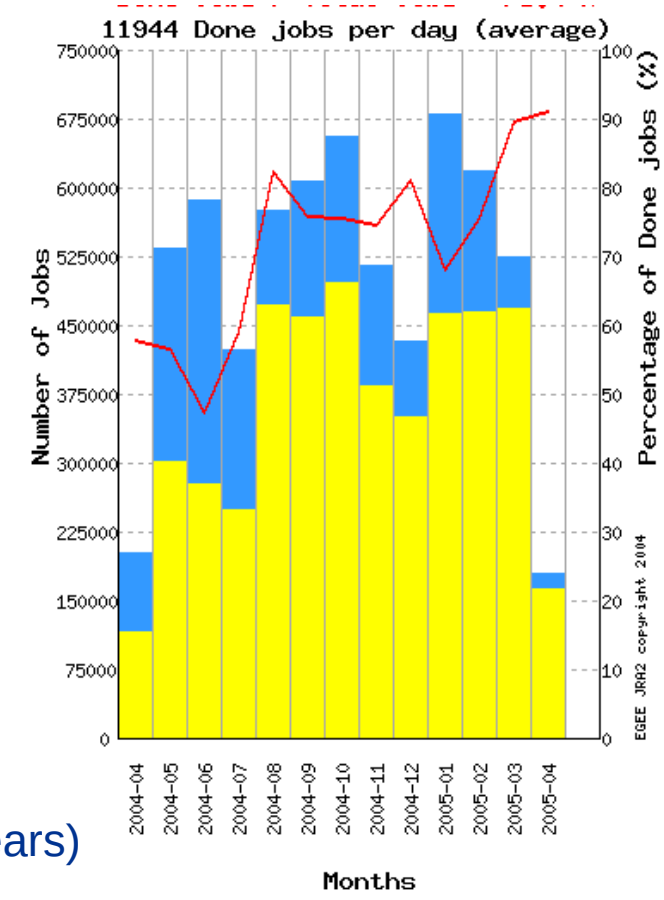
- **Operation of Production Service: real-time display of grid operations**
- **Accounting Information**
- **Selection of Monitoring tools:**
  - GIIS Monitor + Monitor Graphs
  - Sites Functional Tests
  - GOC Data Base
  - Scheduled Downtimes

  - 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: 600
  + Many local VOs, supported by their ROCs

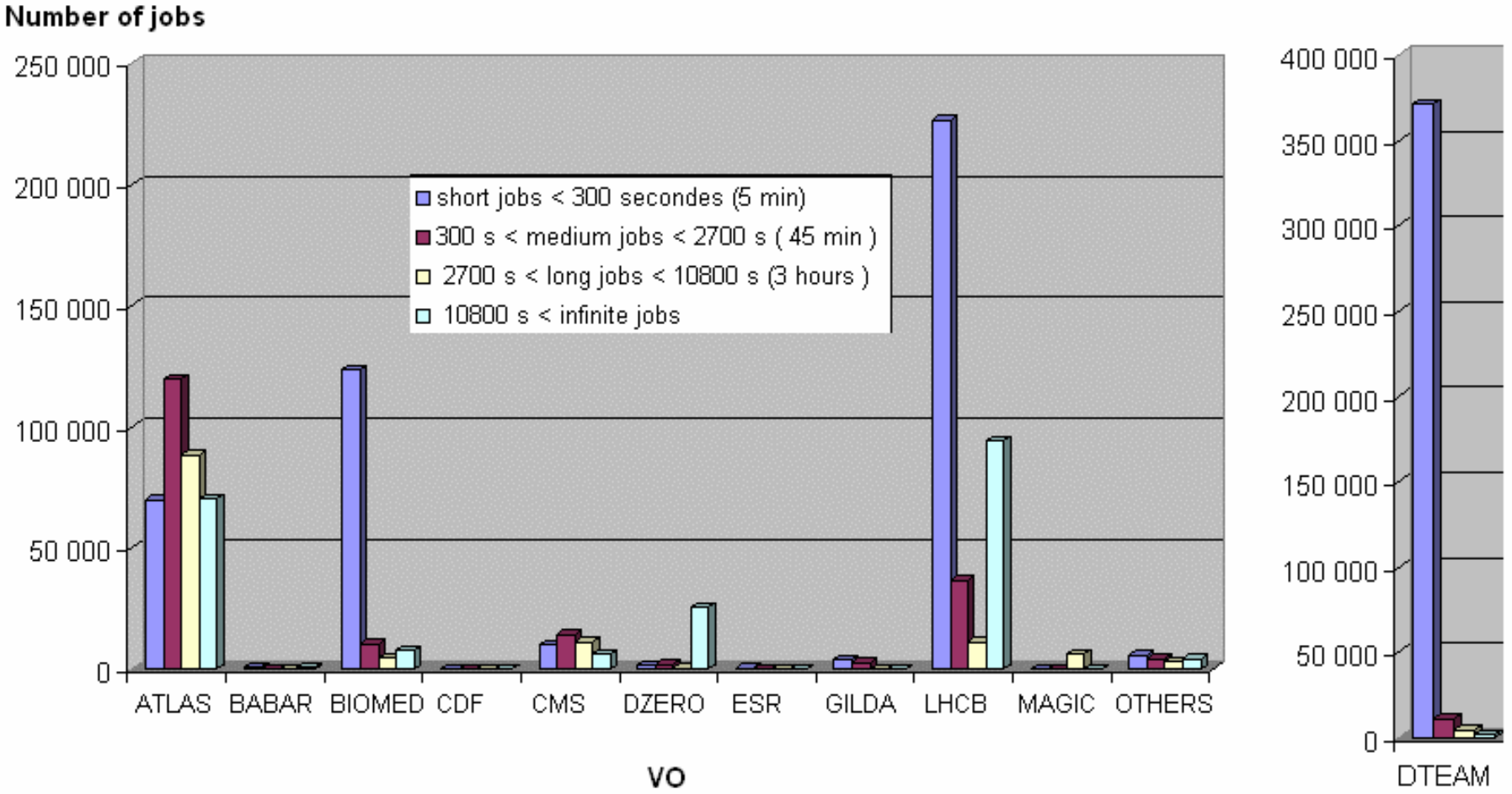
- **Scale of work performed:**
  - LHC Data challenges 2004:
    - >1 M SI2K years of CPU time (~1000 CPU years)
    - 400 TB of data generated, moved and stored
    - 1 VO achieved ~4000 simultaneous jobs (~4 times CERN grid capacity)



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

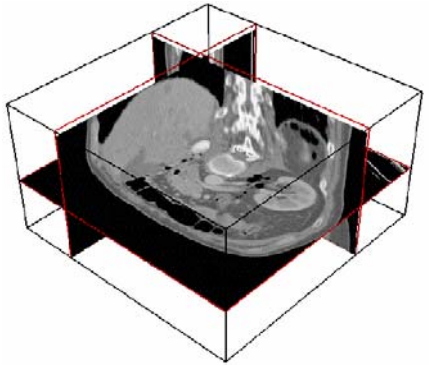
EGEE infrastructure usage

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



EGEE pilot applications (I)

- **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:

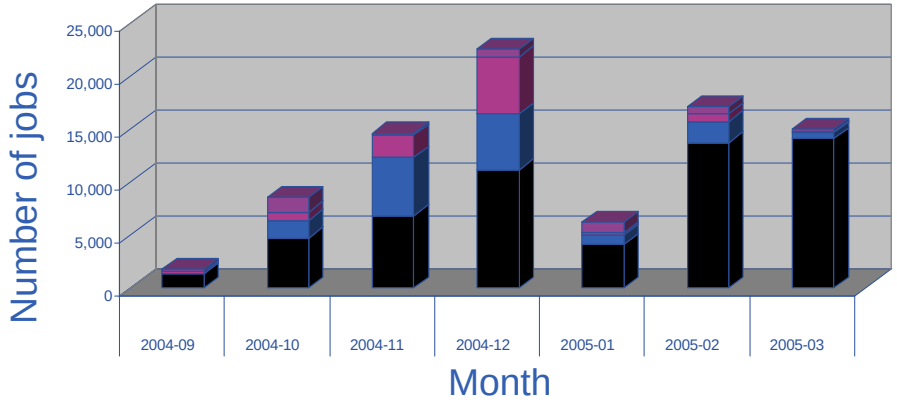
    security & privacy

BioMed Overview

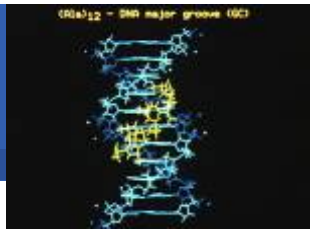
- **Infrastructure**
  - ~2.000 CPUs
  - ~21 TB of disk
  - in 12 countries

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

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



Bioinformatics

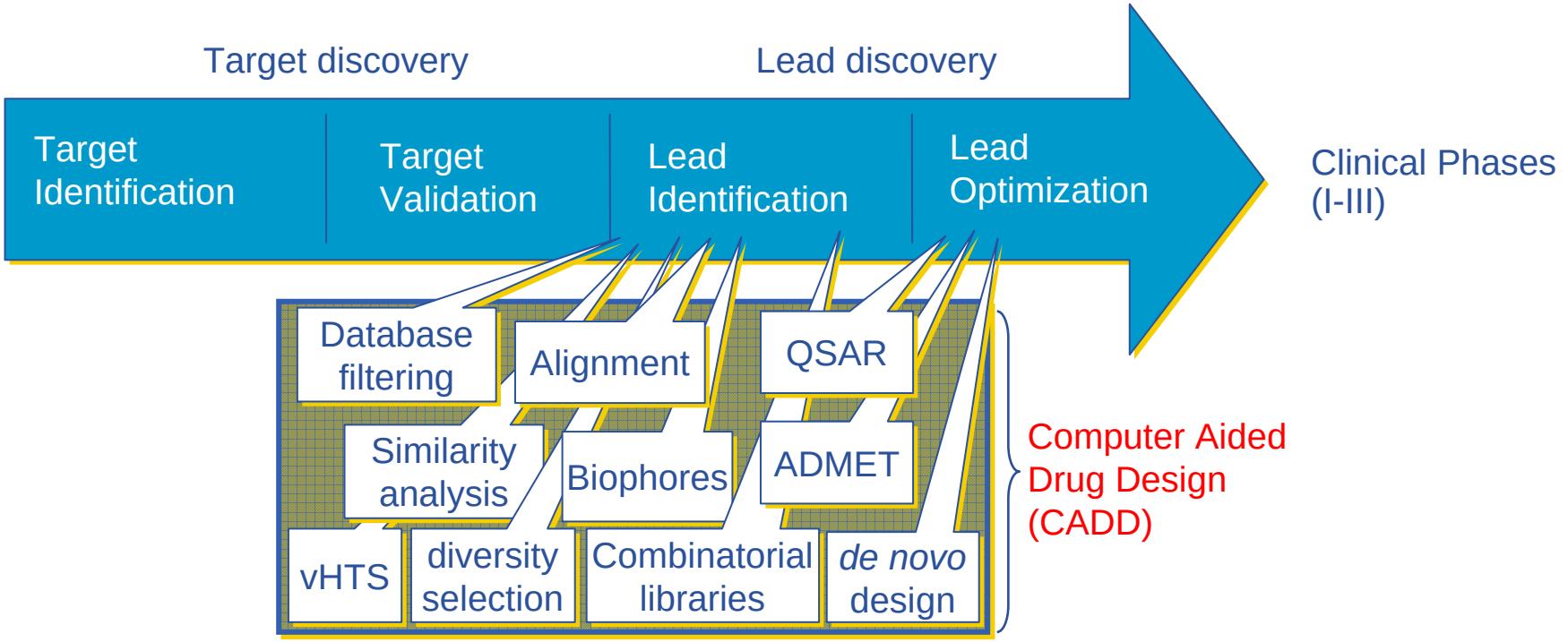


- **GPS@: Grid Protein Sequence Analysis**
  - **Gridified version of NPSA web portal**
    - Offering proteins databases and sequence analysis algorithms to the bioinformaticians (3000 hits per day)
    - Need for large databases and big number of short jobs
  - **Objective**: increased computing power
  - **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

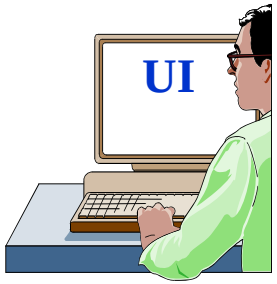
- **Demonstrate the relevance and the impact of the grid approach to address Drug Discovery for neglected diseases**



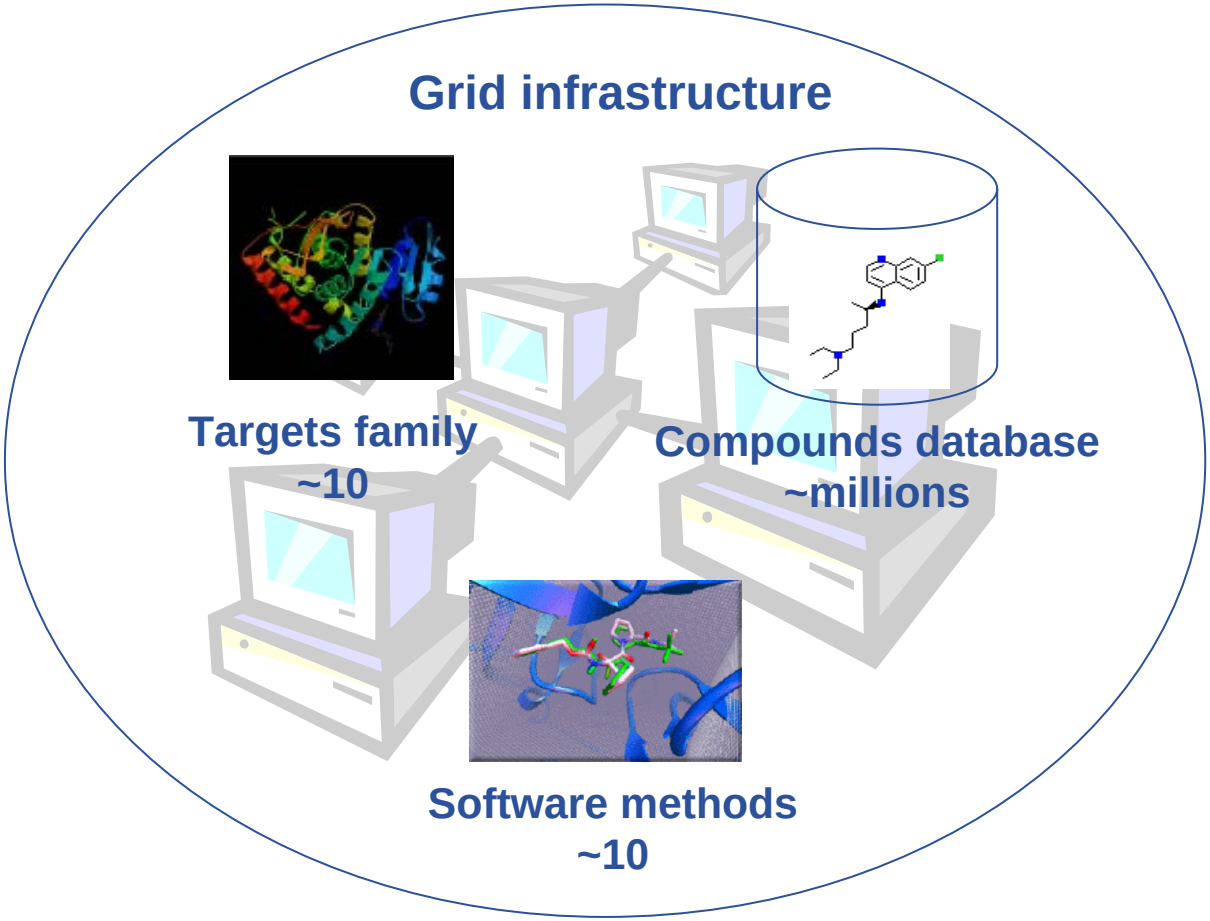
Duration: 12 – 15 years, Costs: 500 - 800 million US $

Docking platform components

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



Parameter / scoring settings



Drug Discovery Data Challenge

- **4 July – 26 August 2005, incl. testing**
  A. 2 weeks using commercial docking software
  B. 3 weeks using free (but slower) docking software

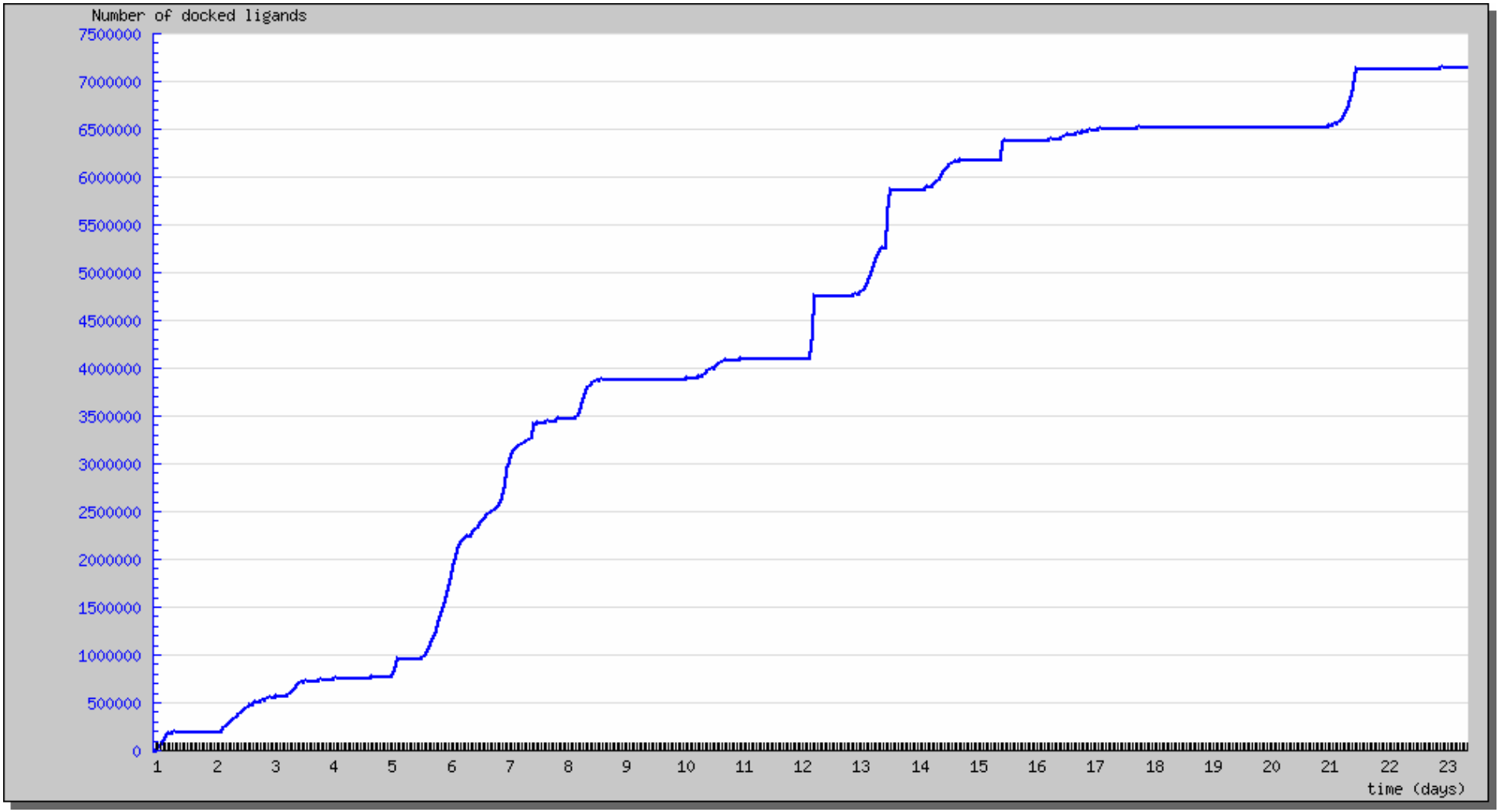
- **Phase A:**
  - 90 packets launched (~ 12900 jobs; 5 to >25 hours each)
  - **~ 20 CPU years** (800 to >1000 CPUs concurrently used)
  - 5800 correct results collected (rest are still running…)
  - file error or failures: 23% → resubmitted
  - 500 GB of data produced

- **Phase B:**
  - 60 packets launched (~30000 jobs; 10 to >25 hours each)
  - **~ 40 CPU years**
  - 1 TB will be produced

- **Final data production: 1,5 TB**

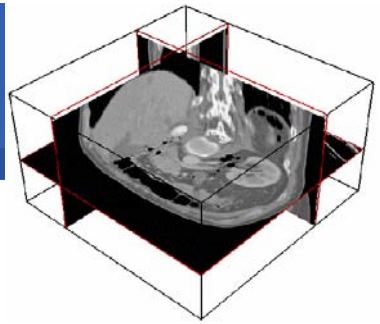
Drug Discovery Data Challenge (II)

- **Number of docked ligands vs. time**



Status 25 July 2005

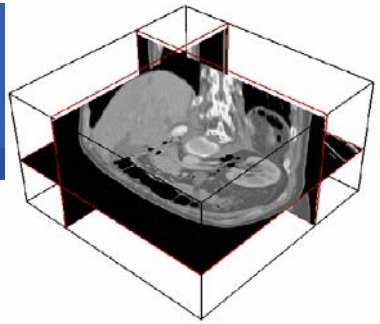
Medical imaging



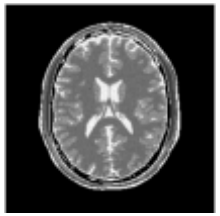
- **GATE**
  - **Radiotherapy planning**
    - Improvement of precision by Monte Carlo simulation
    - Processing of DICOM medical images
  - **Objective**: very short computation time compatible with clinical practice
  - **Status**: development and performance testing
  - **Grid Added Value**: parallelisation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

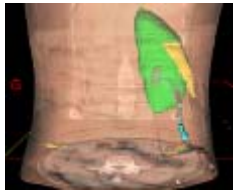


- **SiMRI3D**
  - **3D Magnetic Resonance Image Simulator**
    - MRI physics simulation, parallel implementation
    - Very compute intensive



  - **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



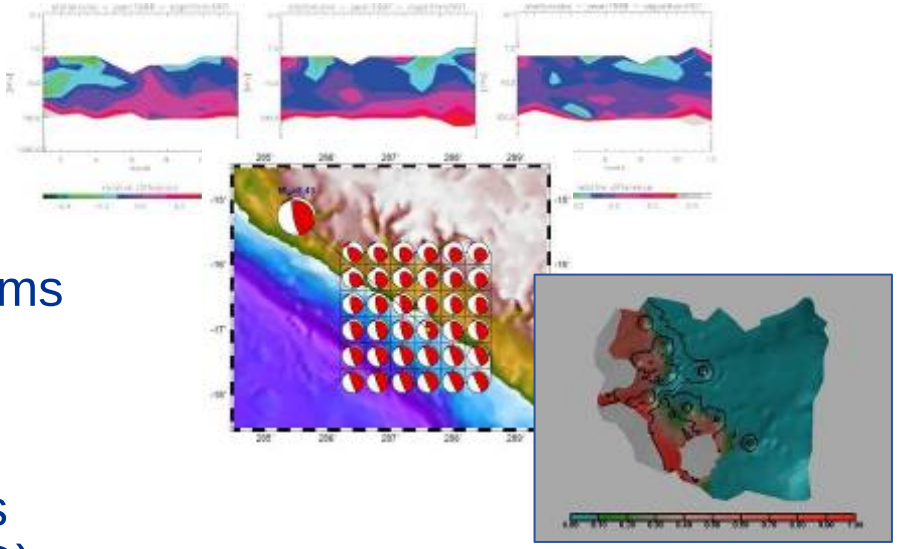
  - **Objectives**: shorten computation time using the grid
    - Interactive reconstruction time: < 2min and scalable
  - **Status**: development of the gridified version being finalized
  - **Grid Added Value**: permanent availability of resources

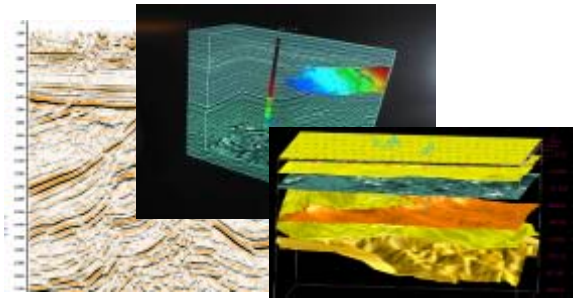
Generic Applications

- **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

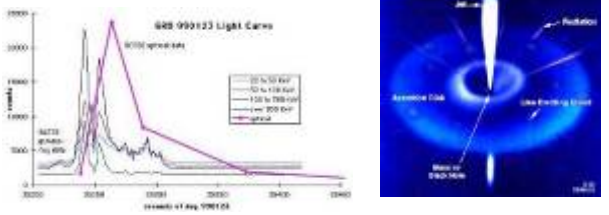




- **A large variety of applications ported on EGEE which incites new users**
- **Interactive Collaboration of the teams around a project**

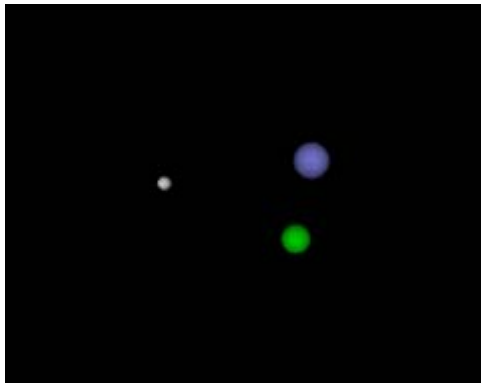
MAGIC

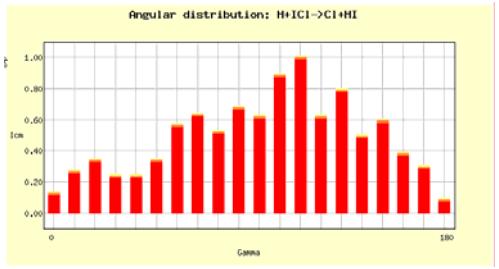
- **Ground based Air Cerenkov Telescope 17 m diameter**
- **Physics Goals:**
  - Origin of VHE Gamma rays
  - Active Galactic Nuclei
  - Supernova Remnants
  - Unidentified EGRET sources
  - Gamma Ray Burst
- **MAGIC II will come 2007**
- **Grid added value**
  - Enable “(e-)scientific” collaboration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

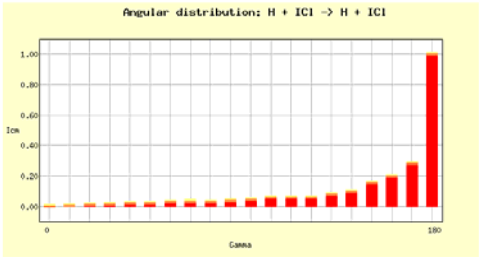


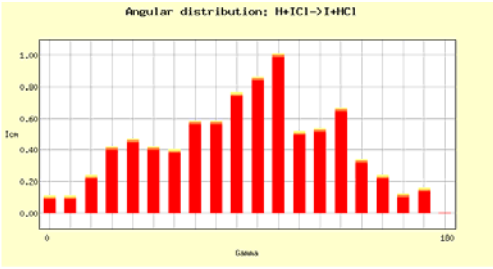
Computational Chemistry

- **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



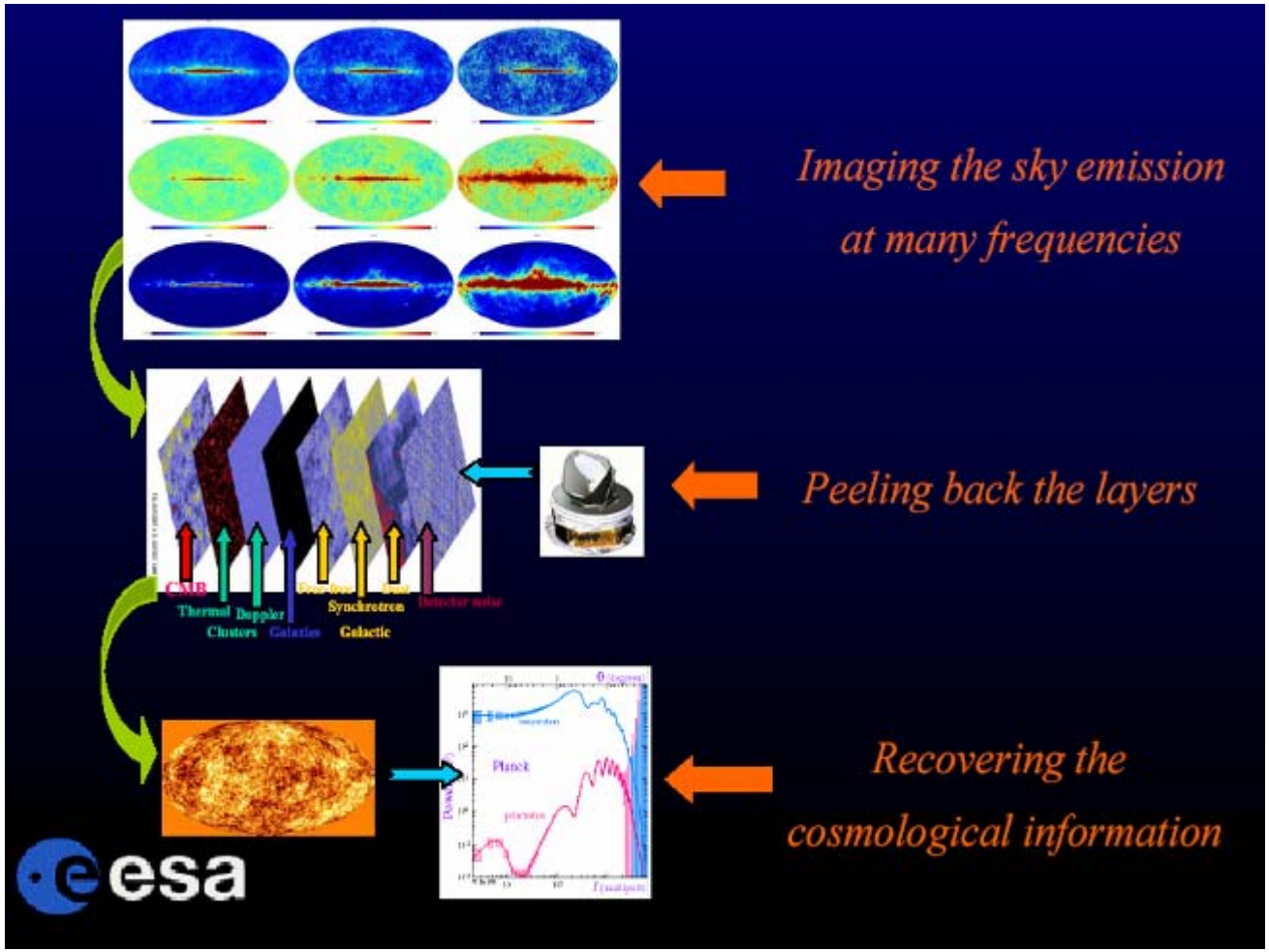






Planck

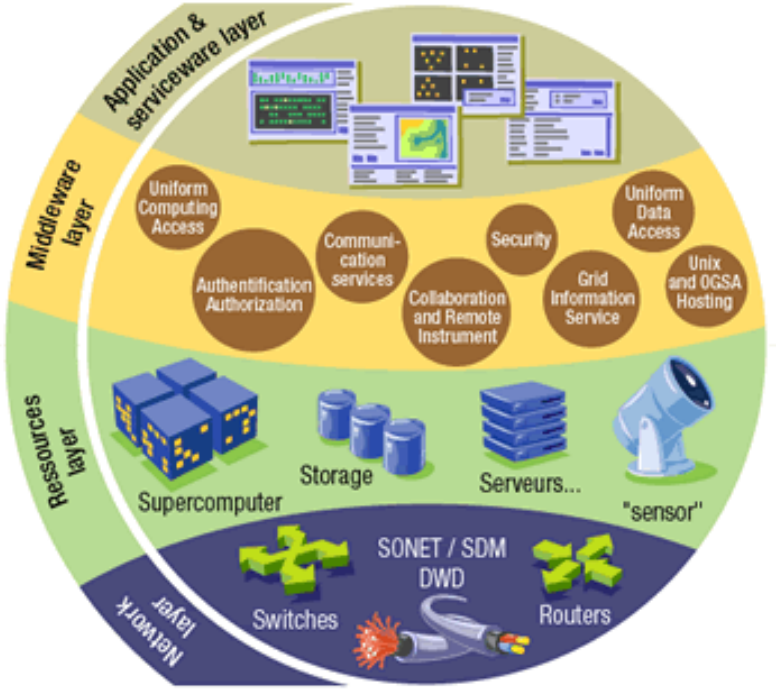
- **On the Grid:**
  > 12 time faster
  (but ~5% failures)
- **Complex data structure**
  → data handling important
- **The Grid as**
  - collaboration tool
  - common user-interface
  - flexible environment
  - new approach to data and S/W sharing



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



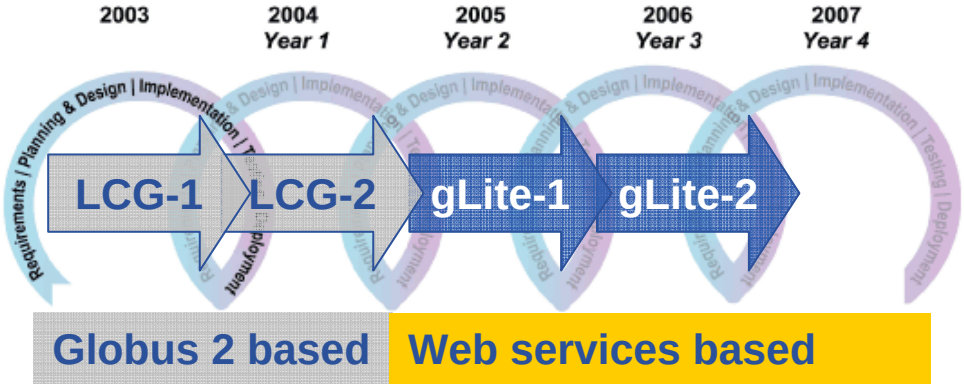
EGEE Middleware gLite

- **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**
  - **see www.gLite.org**
- **Interoperability & Co-existence with deployed infrastructure**
- **Robust: Performance & Fault Tolerance**
- **Service oriented approach**
- **Open source license**

gLite
Lightweight Middleware for Grid Computing

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 140 training events across many countries**
  - >2000 people trained
    induction; application developer; advanced; retreats
  - Material archive 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: 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*

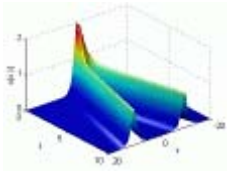
- EGEE as incubator
  - 18 recently submitted EU proposals supported
  - More proposals in next calls and national funding programmes

Related projects under negotiation

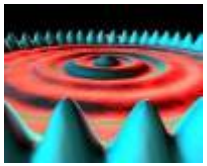
Name	Description	Common partners with EGEE
BalticGrid	EGEE extension to Estonia, Latvia, Lithuania	KTH – PSNC – CERN
EELA	EGEE extension to Brazil, Chile, Cuba, Mexico, Argentina	CSIC – UPV – INFN – CERN – LIP – RED.ES
EUChinaGRID	EGEE extension to China	INFN – CERN – DANTE – GARR – GRNET
EUMedGRID	EGEE 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

Exact budget and partner roles to be confirmed during negotiation

From Phase I to II

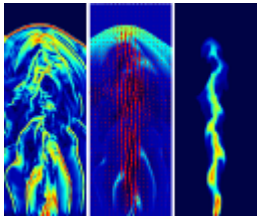


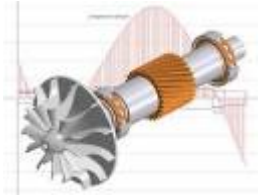
- **From 1st EGEE EU Review in February 2005:**
  - "The reviewers found the overall performance of the project very good."
  - "… remarkable achievement to set up this consortium, to realize appropriate structures to provide the necessary leadership, and to cope with changing requirements."

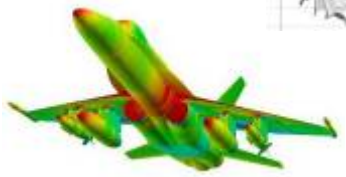


- **EGEE I**
  - Large scale deployment of EGEE infrastructure to deliver production level Grid services with selected number of applications

- **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 I

- 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

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
- **EGEE I and II are always open to further collaboration**
- **The Asia-Pacific region is very important for EGEE and the EU**
  - CKSC is a partner in EGEE, and along with KISTI will form the Korean Federation in EGEE II
- **EGEE is interested in discussing possible future new collaborations**

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

Thanks for the opportunity to present EGEE to all of you and for your kind attention!