# EGEE – Bringing applications to the grid

*Johan Montagnat*
*NA4/Biomed application leader*
*CNRS, France*

**Grid@Work Event, Sophia Antipolis**
**October 10th 2005**

**www.eu-egee.org**

Information Society

# Objectives of EGEE for application deployment (NA4 activity)

- To identify through the dissemination partners and a well defined integration process a portfolio of early user applications from a broad range of application sectors from academia, industry and commerce

- To support development and production use of all of these applications on the EGEE infrastructure and thereby establish a strong user base on which to build a broad EGEE user community

- To initially focus on two well-defined pilot application areas, Particle Physics and Biomedicine

# NA4 activities

- **Application deployment**
  - act as a liaison between the applications and the operational infrastructure
  - ensure that the minimum services for all supported applications are provided.

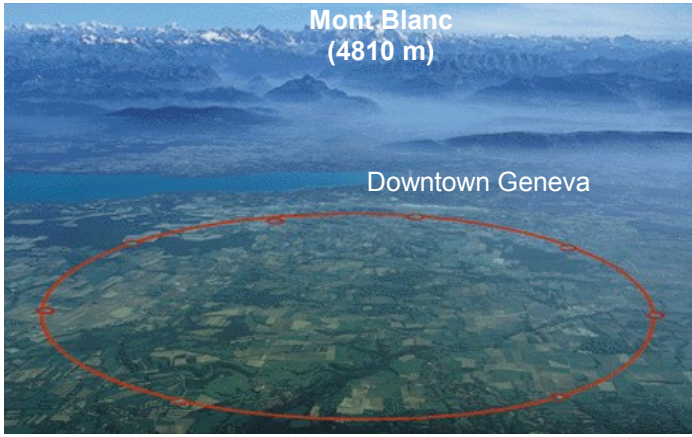
- **Troubleshooting/testing**
  - help applications resolve problems with deployment
  - filter problems before they are submitted to operations support (SA1)
  - proactively test the infrastructure (from an application perspective) to identify and resolve problems before they affect production use

# NA4 activities

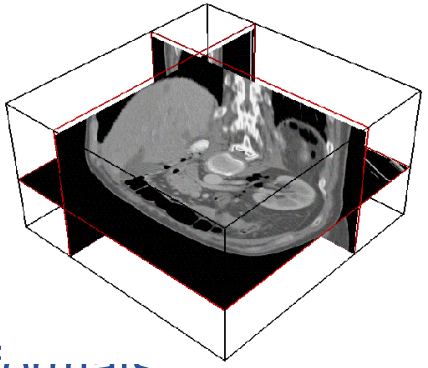
- **Consulting, porting of applications**
  - support for applications to port their software to the grid environment
  - recurrent support for existing applications as new services appear in the grid middleware and the applications evolve
- **Evolution of pilot applications:**
  - to exploit the required grid baseline services deployed in EGEE
  - this activity should be designed together with the applications
- **Validation of the EGEE infrastructure**
  - to guarantee the production and availability of the data for analysis by the large and distributed scientific communities
  - follow carefully functionality and performance issues of the applications on the production service

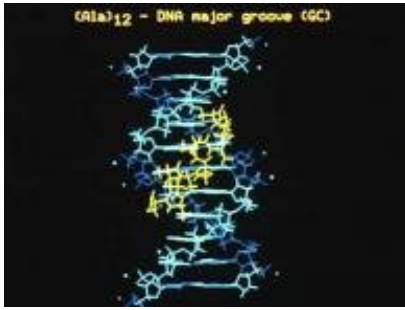
# EGEE pilot applications

- **High-Energy Physics (HEP)**
  - Provides computing infrastructure (LCG)
    - for experiments at CERN in Geneva
  - 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 & access to data in many formats





# "Generic" applications

- **New Communities identification**
  - Ease access to the infrastructure
  - Provide support
  - Help in application gridification
- **Application areas**
  - Earth Sciences
  - Geology (petrol industry)
  - Computational Chemistry
  - Astrophysics
  - Grid search engines

# EGEE Industry Forum

- **Objectives:**
  - To promote and disseminate Grid concepts towards industry and service groups
  - To raise the awareness of EGEE within industry
  - To encourage businesses to participate in the project

- **Members:** **interested companies having activities in Europe**

- **Activities:**
  - Organisation of a meeting twice a year
  - Quarterly newsletter
  - Participation to EGEE working groups (EGAAP, Project Technical Forum, EGEE Phase 2, Security group)
  - Internal Working groups
    - Technical aspects of Grid
    - Business models and economical aspects

# VO: Enabling users groups

- **Scientific communities are gathers in Virtual Organizations (VO)**
  - 4 High Energy Physics VOs (different Large Hadron Collider experiments)
  - Biomed VO
  - Earth Sciences VO
  - ...
- **People in a VO share**
  - Data (in LCG2, finer authorization grain to come)
  - Infrastructure access rights
  - Core services specific to the VO
  - Administration

# Evolution of the number of users

| Virtual Organization Name | Discipline | Number of users in January 2005 | Number of users in September 2005 |
|---|---|---|---|
| ALICE | HEP (LHC experiment) | 27 | 50 |
| ATLAS | HEP (LHC experiment) | 203 | 400 |
| CMS | HEP (LHC experiment) | 161 | 350 |
| LHCb | HEP (LHC experiment) | 41 | 58 |
| ESR | Earth Sciences | 18 | 33 |
| Biomed | Medical/bioinformatics | 33 | 67 |
| CompChem | Chemistry | 9 | 5 |
| Magic | Astronomy | 5 | 10 |
| EGEODE | GeoPhysics | 0 | 3 |
| Planck | Astrophysics | 0 | 5 |
| *TOTAL* | | *497* | *981* |

# User information & support

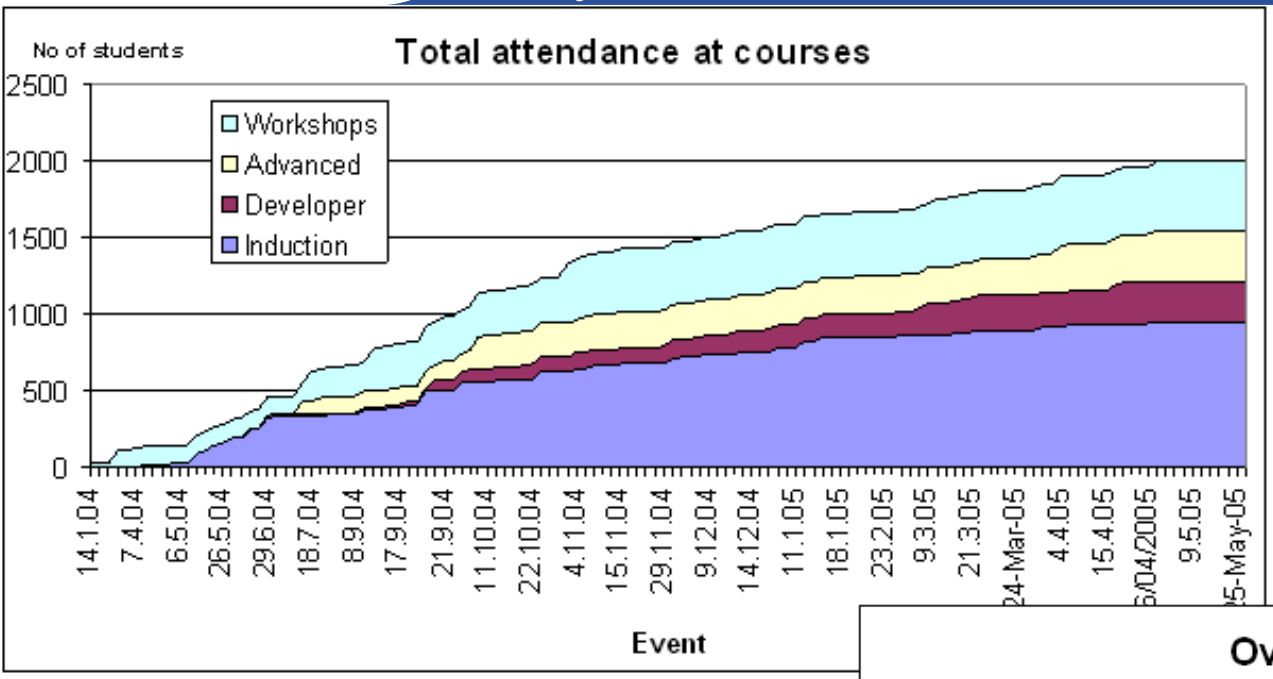
- **More than 140 training events across many countries**
  - >2000 people trained
    induction; application developer; advanced; retreats
  - Material archive online with >200 presentations

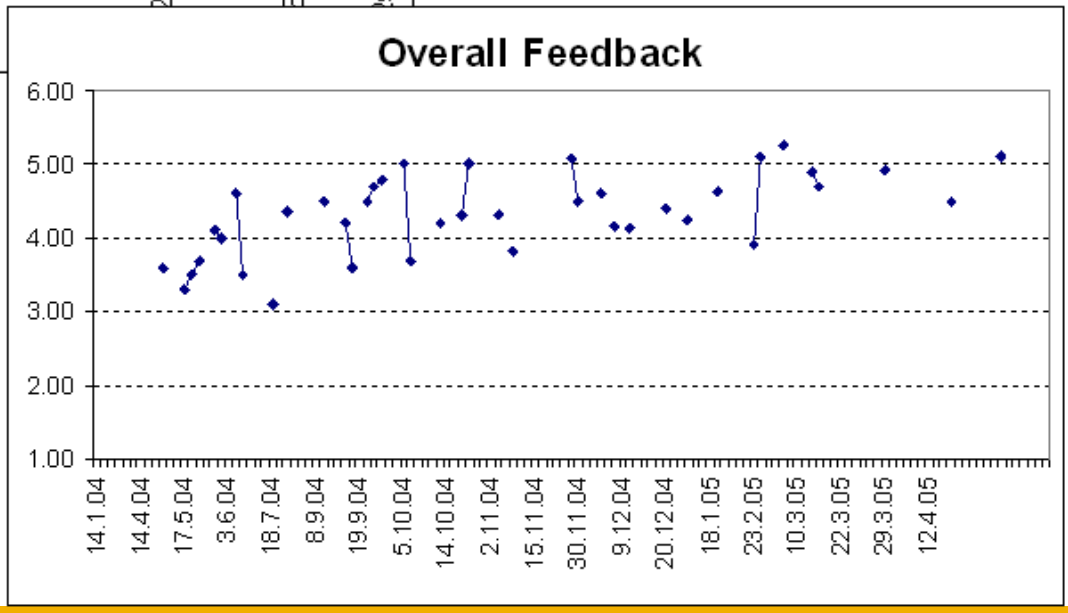
- **Public and technical websites constantly evolving to expand information available and keep it up to date**

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

- **Pisa: 4th project conference 24-28 October '05**
  - Registration open: http://public.eu-egee.org/conferences/4th/

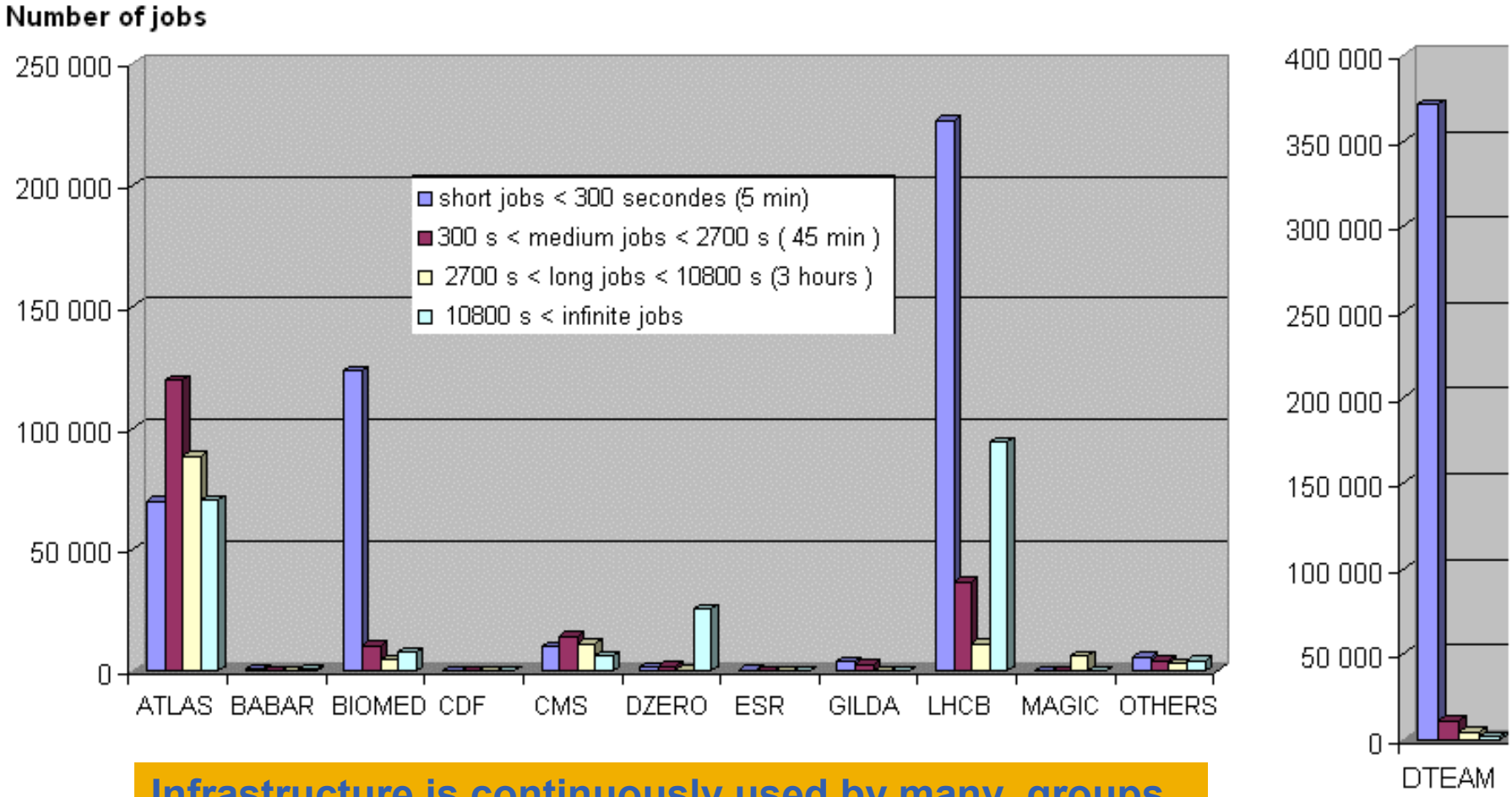
# Training achievements



# EGEE infrastructure usage

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




**Infrastructure is continuously used by many groups**

# EGEE infrastructures

- **LG2: Production infrastructure**
  - LCG2 middleware
  - Batch-oriented, Command Line Interface
  - Resource brokering, Workload management, File management...
  - Large scale infrastructure
    - 14000 CPUs (130 centers)
    - 5 PB storage space
- **GILDA: Induction infrastructure**
  - LCG2 / gLite middleware
  - Web interface (GENIUS)
  - Smaller scale (~200 CPUs)
- **PPS: Next generation middleware testing infrastructure**
  - gLite middleware
  - Improved functionalities (File Transfer, Security, Group of jobs...)
  - Testing infrastructure

# GILDA, an infrastructure for dissemination and demonstration

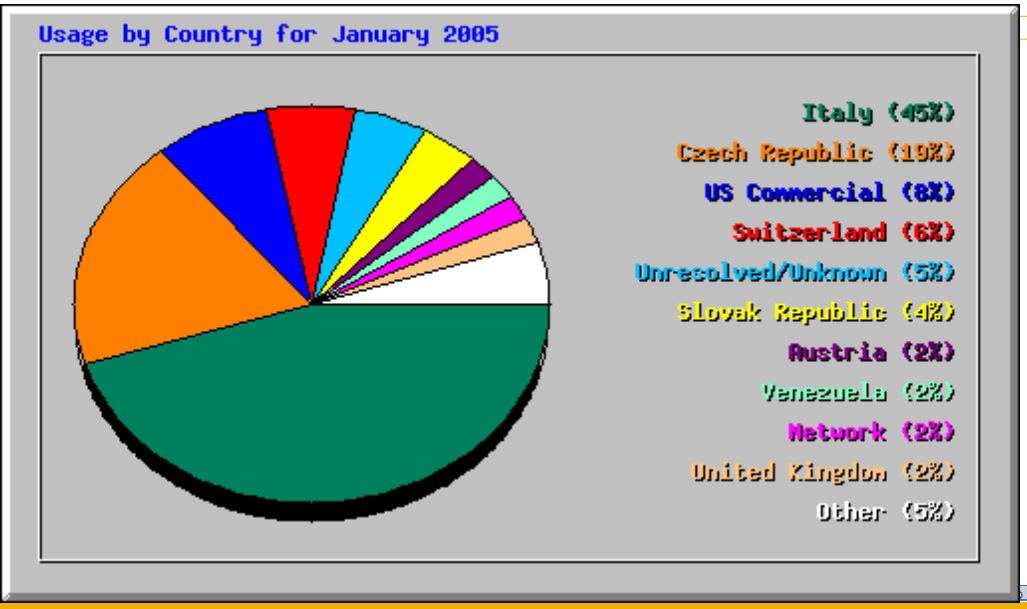
- **Goals**
  - Demonstration of grid operation for tutorials and outreach
  - Initial deployment of new applications for testing purposes

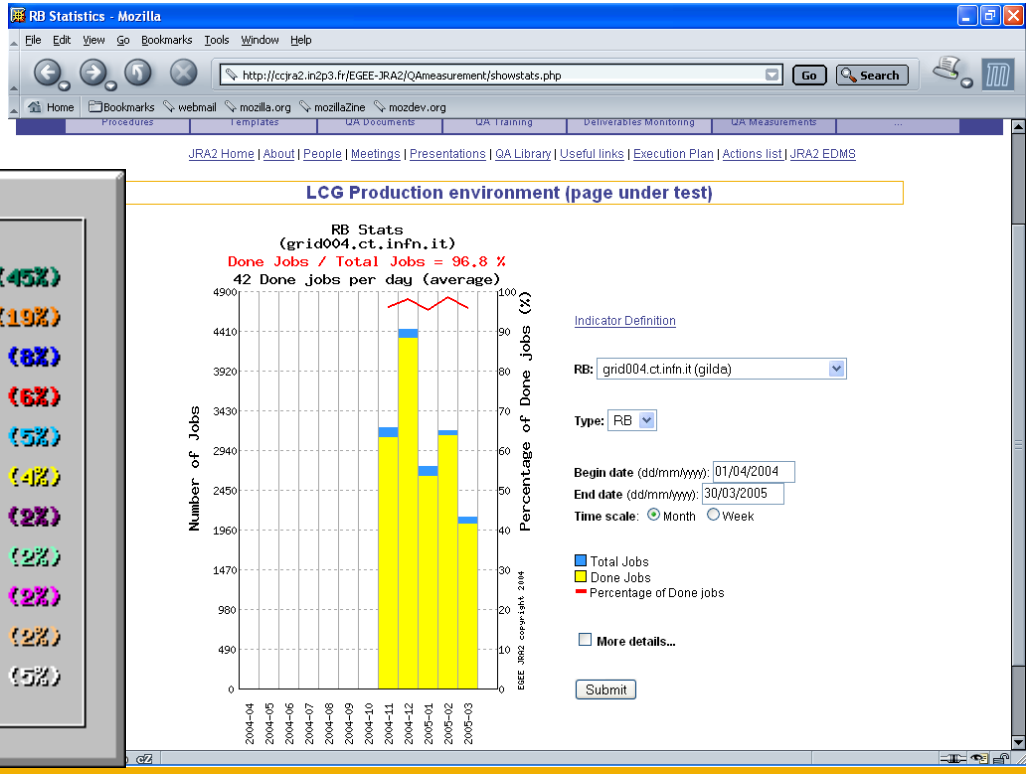
- **Key features**
  - Initiative of the INFN Grid Project using LCG-2 middleware
  - On request, anyone can quickly receive a grid certificate and a VO membership allowing them to use the infrastructure for 2 weeks
  - Certificate expires after two weeks but can be renewed
  - Use of friendly interface: Genius grid portal

- **Very important for the first steps of new user communities on to the grid infrastructure**

# GILDA summary numbers

- 15 sites in 3 continents
- > 2000 certificates issued, 15% renewed at least once
- > 50 tutorials and demos performed in 15 months
- > 50 jobs/day on the average
- Job success rate above 80%
- > 750,000 hits (> 40,000 visits) on (of) the web site from 10's of different countries
- > 0.5 TB of videos and UI's downloaded from the web site



# The GILDA Tutorials/Demonstrations (1/2) (https://gilda.ct.infn.it/tutorials.html)

## 2004

Edinburgh, 7 April 2004, slides, pictures
Tunis, 22-23 April 2004, pictures
Edinburgh, 26-28 April 2004, slides, pictures
CERN, 17-19 May 2004, pictures
Catania, 24-25 May 2004, home page, pictures
Dubna, 29 June - 2 July 2004, agenda
Edinburgh, 6 July 2004, home page
Catania, 14-16 July 2004, home page, pictures
Vico Equense, 19 July 2004, slides, pictures
Vico Equense, 6-10 September 2004, home page
Catania, 4-8 October 2004, home page, agenda
Vilnius, 5-6 October 2004, agenda
London, 6 October 2004
Madrid, 6-7 October 2004, agenda
Heidelberg, 11-14 October 2004
CERN, 16 October 2004
Prague, 26 october 2004, home page
Warsaw, 4-6 November 2004, home page, agenda
Lyon, 9-10 November 2004, agenda
The Hague, 15-17 November 2004, pictures
Merida, 15-20 November 2004, home page, agenda, slides, pictures
Tunis, 20 November 2004
Rio de Janeiro, 22-23 November 2004, home page, agenda, pictures
The Hague, 24 November 2004, agenda
CERN, 29-30 November 2004, agenda
Kosice, 30 November - 1 December 2004, agenda
Tunis, 6-7 December 2004
Bochum, 7-10 December 2004, home page, agenda
Edinburgh, 8 December 2004, home page
Istanbul, 9-10 December 2004, agenda, slides, pictures
Shanghai, 9-10 December 2004, agenda
Aurillac, 13-14 December 2004
Prague, 16 December 2004, home page, pictures
Tel Aviv, 22-23 December 2004, agenda, pictures

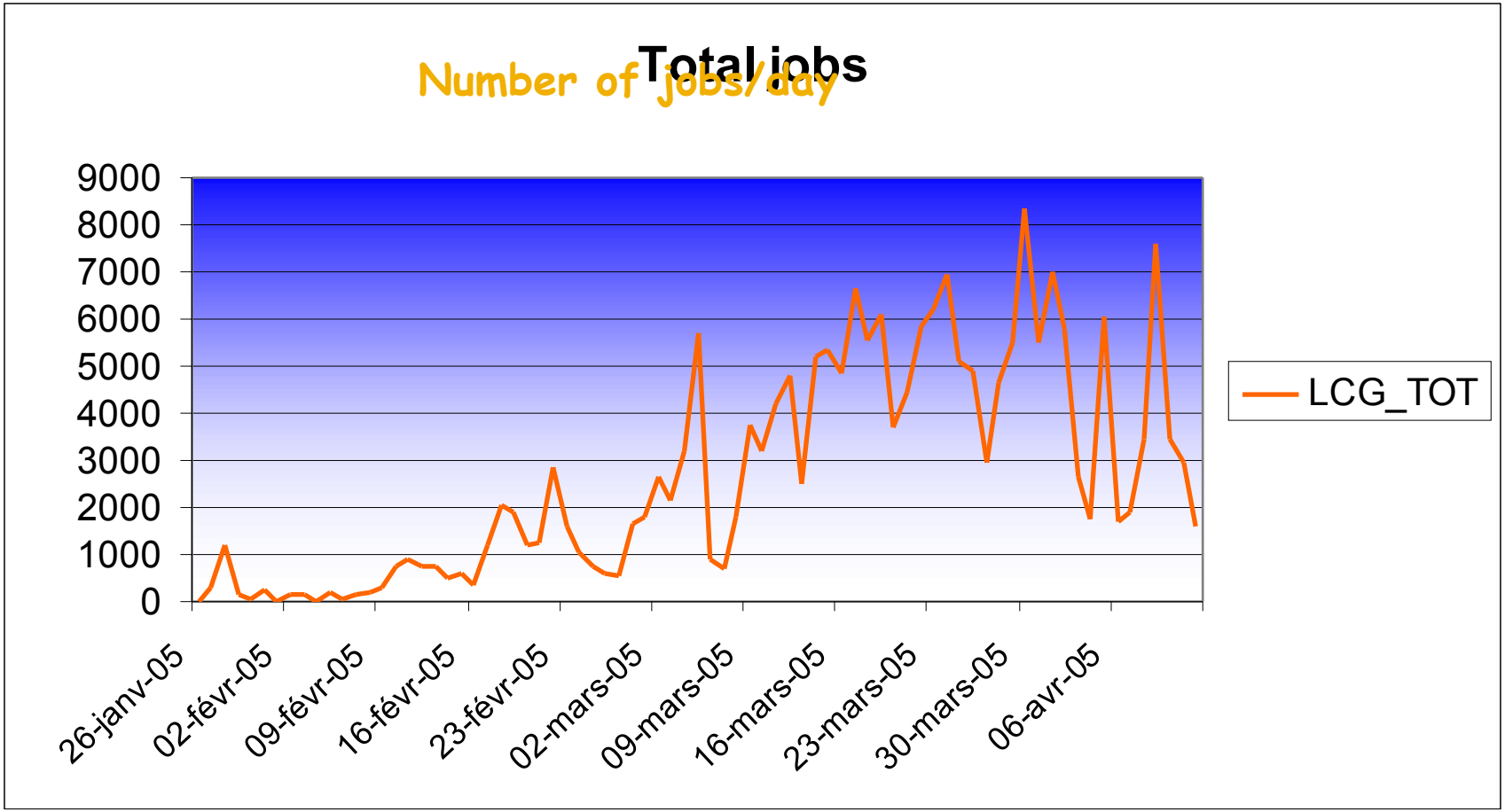
## 2005

CERN, 13 January 2005, agenda
Torino, 18-19 January 2005, home page, agenda
CERN, 20 January 2005, agenda
CERN, 2-4 February 2005, agenda
Roma, 3 February 2005, home page, agenda, pictures
Sydney, 3-4 February 2005, home page
CERN, 9-11 February 2005, agenda
Amsterdam, 14-16 February 2005, home page
Trento, 23-25 February 2005, home page, agenda
Amsterdam, 28 February - 1 March 2005, home page
Julich, 9 March 2005,
Clermont-Ferrand, 9-31 March 2005, agenda
Vienna, March-August 2005
Hamburg, 23-24 March 2005, home page, agenda
Ula-Merida, 31 March-1 April 2005, agenda
Zilina, 4 April 2005, home page and agenda
Edinburgh, 9-13 May 2005, home page and agenda
St.Augustin,25 May 2005, home page and agenda
Catania, 13-15 June 2005, home page, agenda, pictures
Valencia, 14-16 June 2005, home page, agenda
Lyon, 17 June 2005, home page and agenda
Bratislava, 27-30 June 2005, agenda
Forschungszentrum Karlsruhe, 08 July 2005, home page and agenda
Vico Equense, 10-22 July 2005, home page, agenda
Budapest, 11-16 July 2005, home page, agenda
Clermont-Ferrand, 25-27 July 2005, home page, agenda
Madrid, 26-27 July 2005, home page, agenda
Swansea, 06 August 2005, home page and agenda
Taipei, 22-23 August 2005, home page and agenda
Tokyo, 25-26 August 2005, home page and agenda
Seoul, 29-30 August 2005, home page and agenda
Saint-Malo, 12-15 September 2005, home page, agenda

# Applications: High Energy Physics

- **Grid is an essential ingredient for the experiments' success**
  - Pioneering work in distributed computing
  - Involvement in many projects worldwide and users of several grids
  - Very large and distributed scientific community
    - 4 experiments: ALICE, ATLAS, CMS and LHCb
    - ATLAS: ~2000 physicists in over 30 countries
      http://www.cerncourier.com/main/article/39/1/12
  - Grid is an essential ingredient for the experiments' success
- **Production infrastructure (LCG/EGEE)**
  - Intensive usage: data challenges (first and largest users)
    - E.g. LHCb – 3500+ concurrent jobs for long periods (many months)
  - Many issues discovered, but significant work could be done:
    - >1 M SI2K years of cpu time (~1000 CPU years)
    - 400 TB of data generated, moved (replicated) and stored
    - 10k simultaneous jobs (~8 times CERN grid capacity)
- **HEP role in application development and middleware testing**
  - Large effort on the 4 LHC experiments' prototypes (ARDA project)
  - Early usage/feedback for the gLite middleware

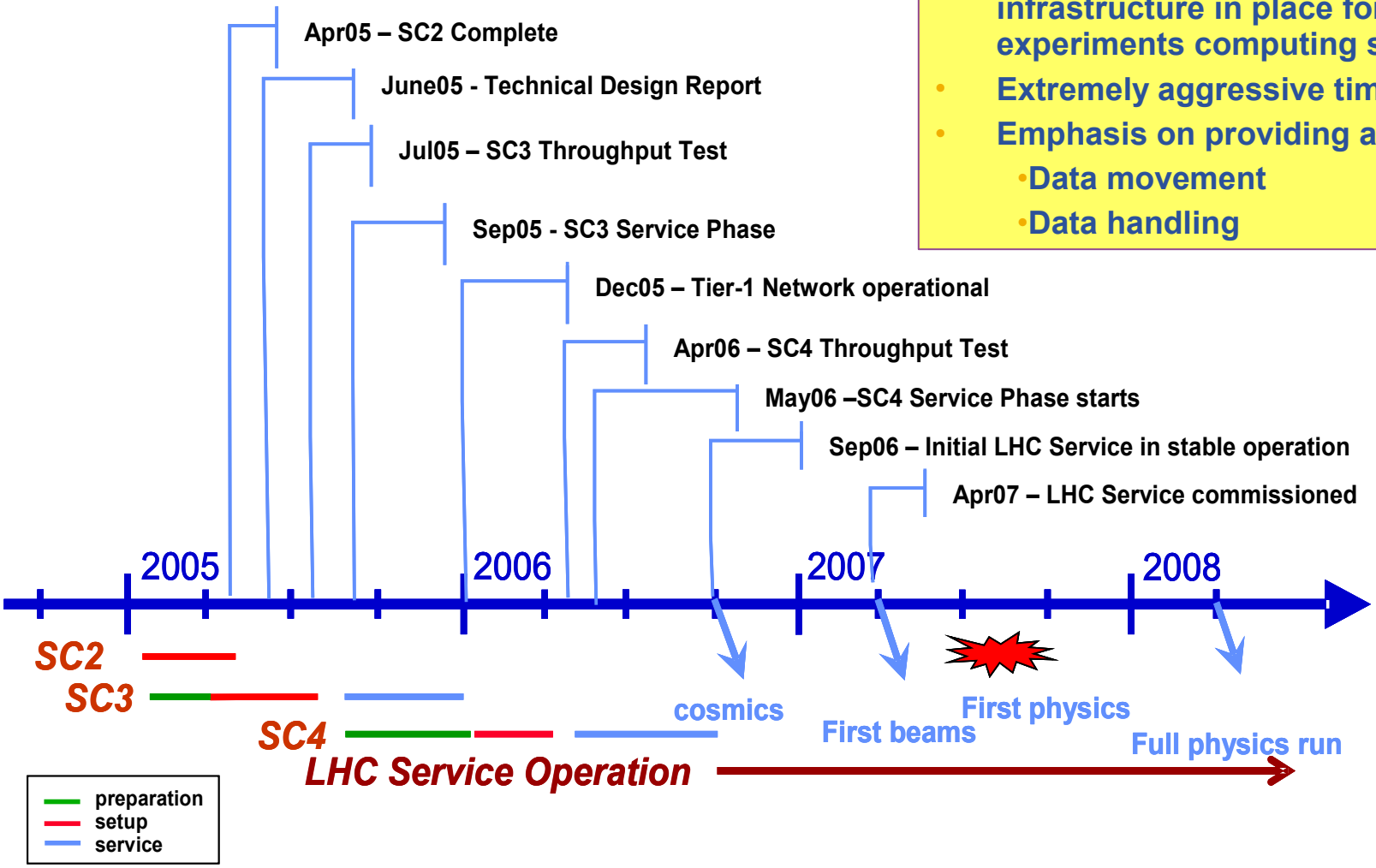
# Recent ATLAS work




- ATLAS jobs in EGEE/LCG-2 in 2005
  - In latest period up to 8K jobs/day (peak over 10k concurrent long jobs)

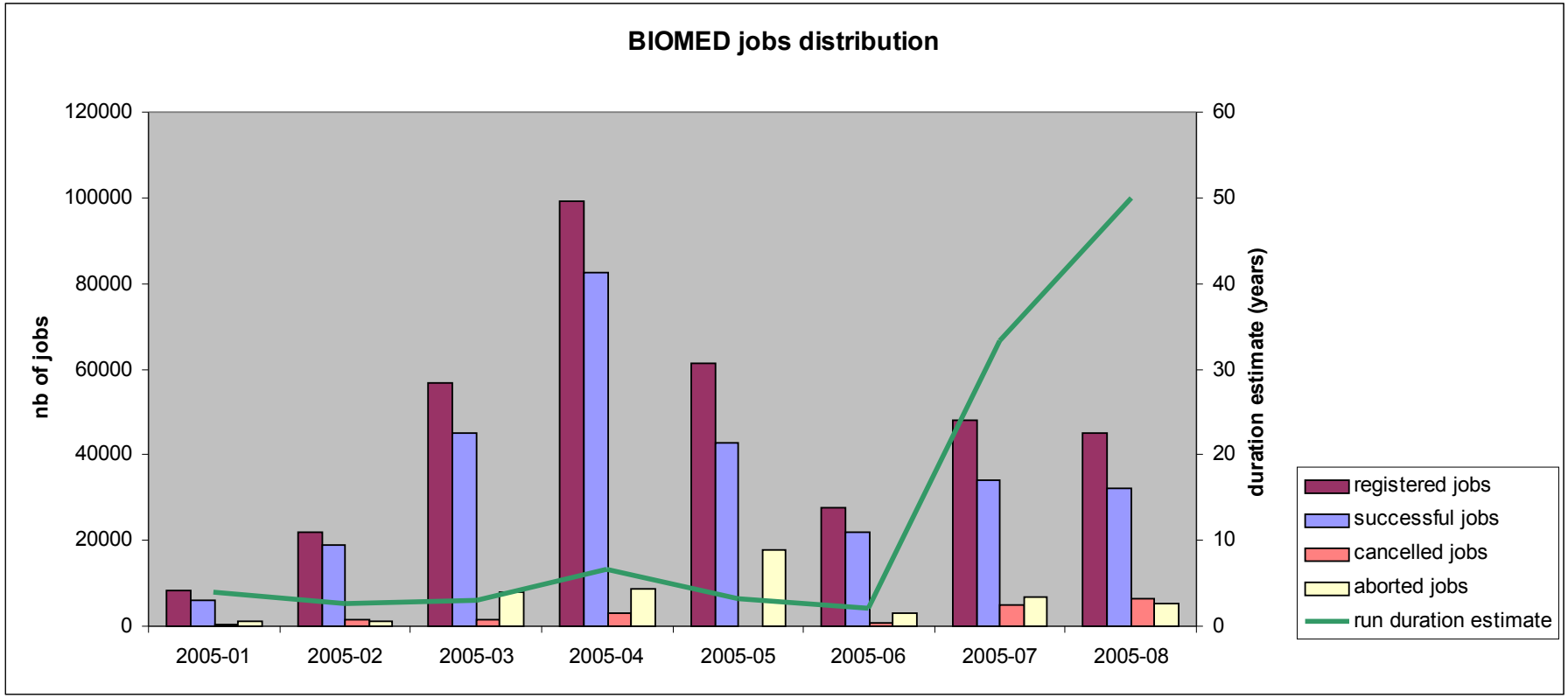
# How to prepare for LHC (1): LCG Service Challenges

- LHC starts in 2007
- Ramp-up with series of service challenges to ensure key services & infrastructure in place for the experiments computing systems
- Extremely aggressive timescale
- Emphasis on providing a service
  - Data movement
  - Data handling



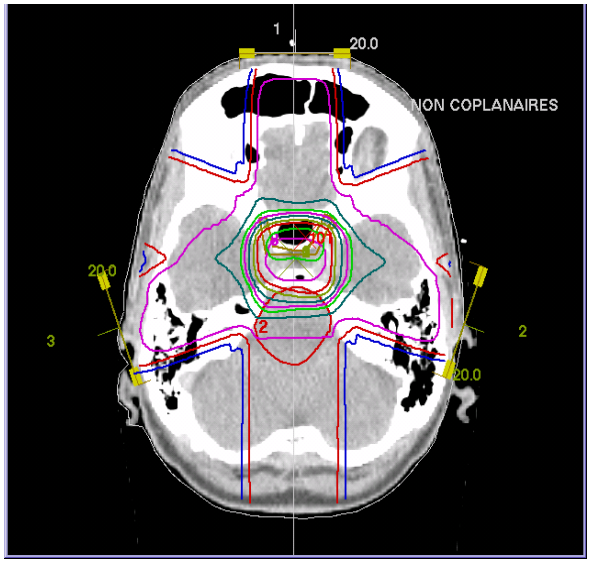
# Biomed Virtual Organisation

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

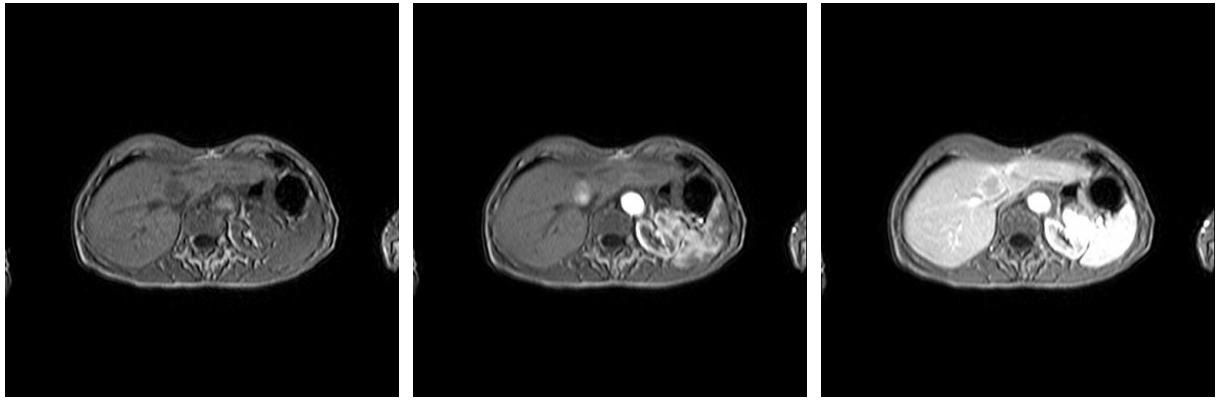


# Medical image processing

- GATE: Radiotherapy planning
  - Monte Carlo simulation
  - Parallel execution on different seeds

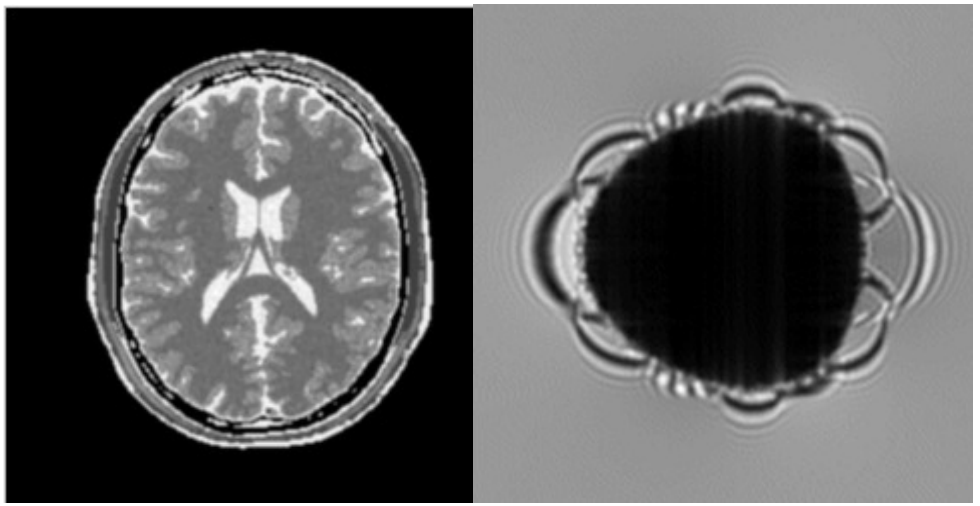



- Pharmacokinetics: contrast agent diffusion study
  - Medical images registration
  - Distribution of registration pairs

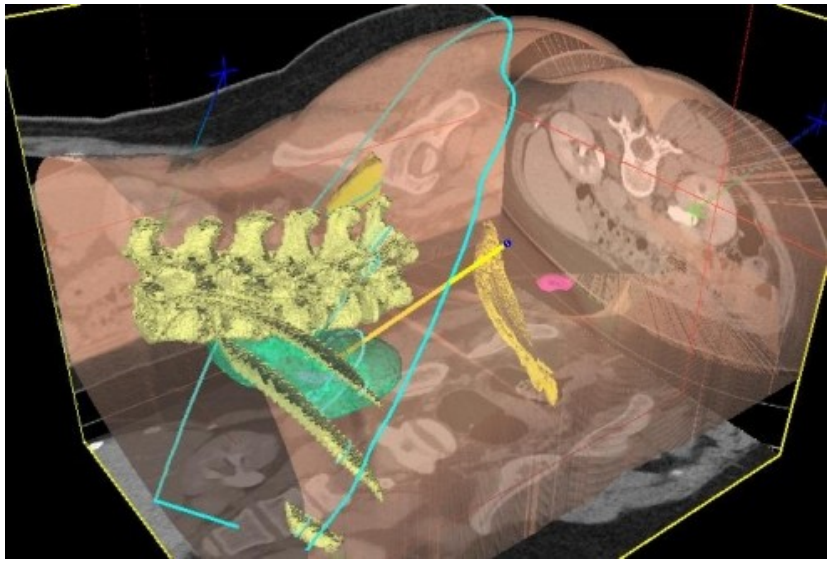


# Medical image processing

- SiMRI3D MRI simulation
  - Magnetic Resonance physics simulation (Bloch's equation)
  - Parallel processing (MPI)

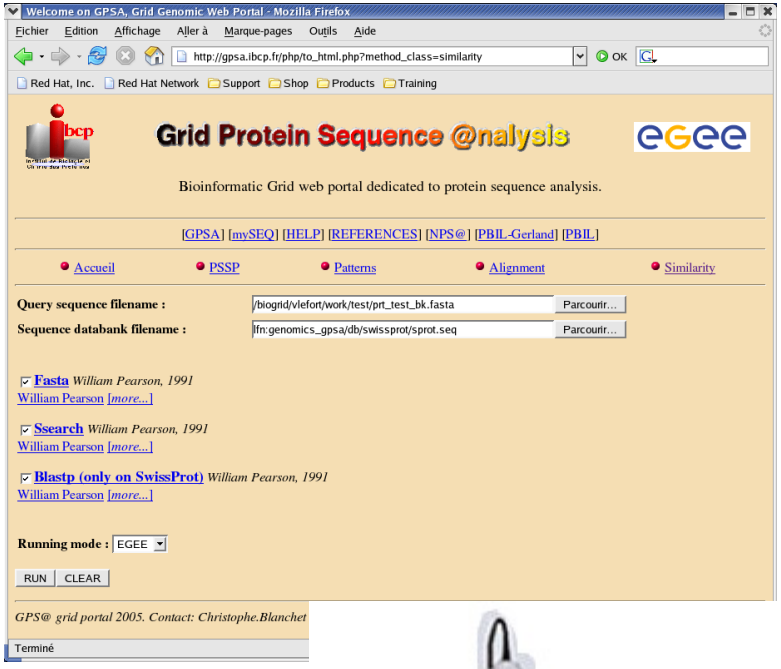


- gPTM3D: Radiological images segmentation tool
  - Deformable-contour based segmentation
  - Interactivity through agent-based scheduling



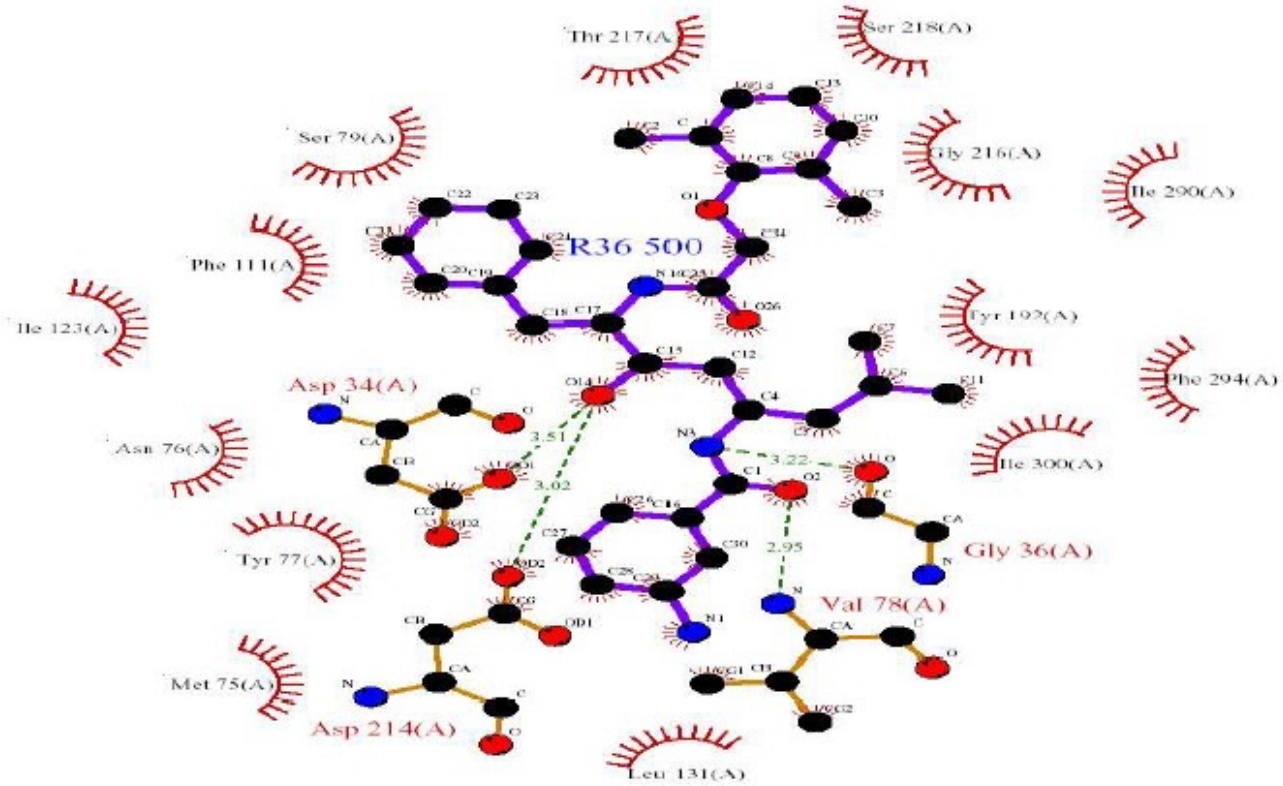
# Bioinformatics

- GPS@: bioinformatics portal
  - http://gpsa.ibcp.fr/ web portal
  - Existing (but overloaded NPSA portal)
  - Tens of bioinformatics legacy code
  - Thousands of potential users
  - Large input databases




- Electron-microscopic image reconstruction
  - Image filtering and noise reduction
  - 3D structure analysis

- Molecular docking
  - Target protein from Malaria genome
  - Testing geometrical docking of ligand databases (potential drugs) against targets

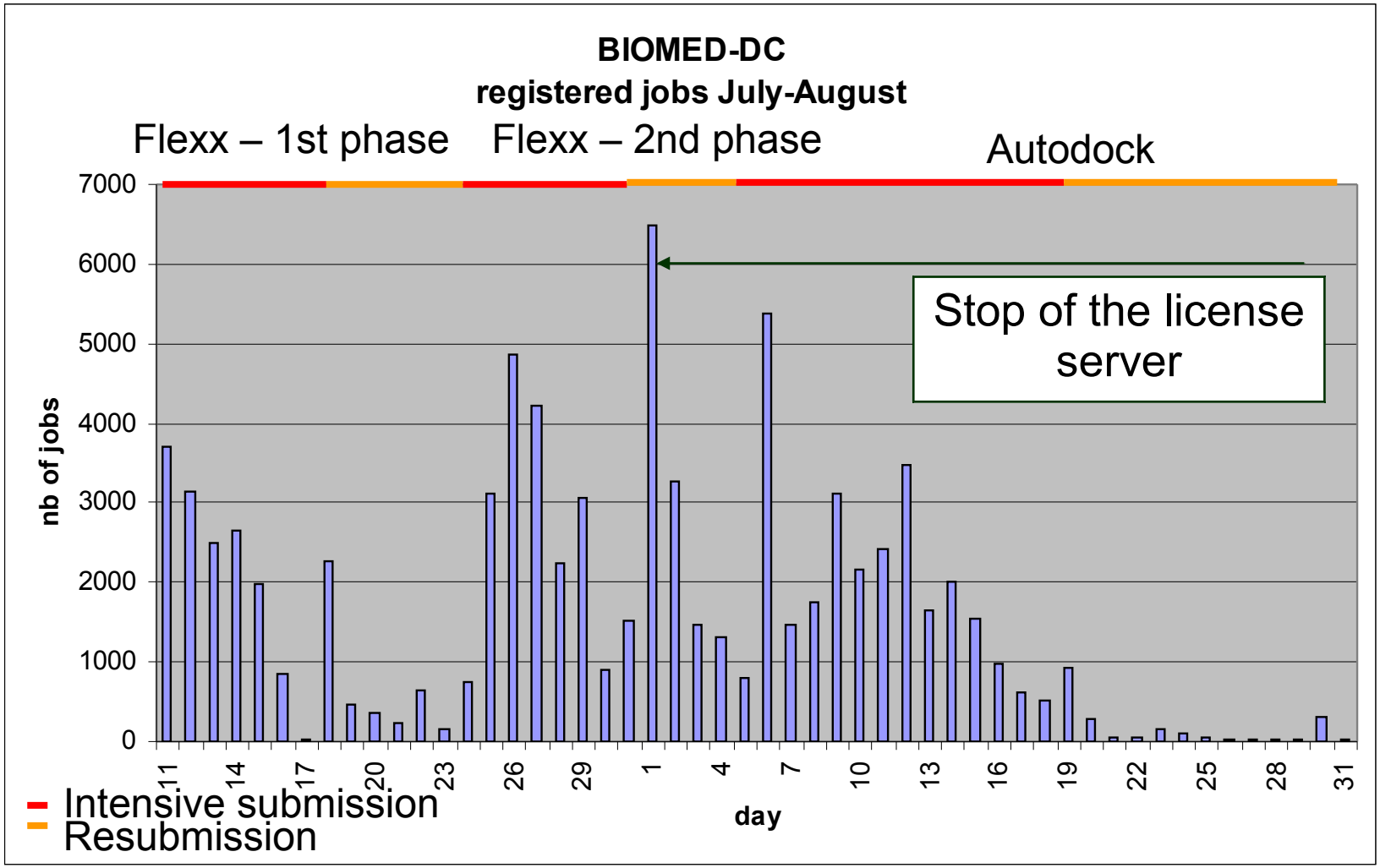


# WISDOM elements

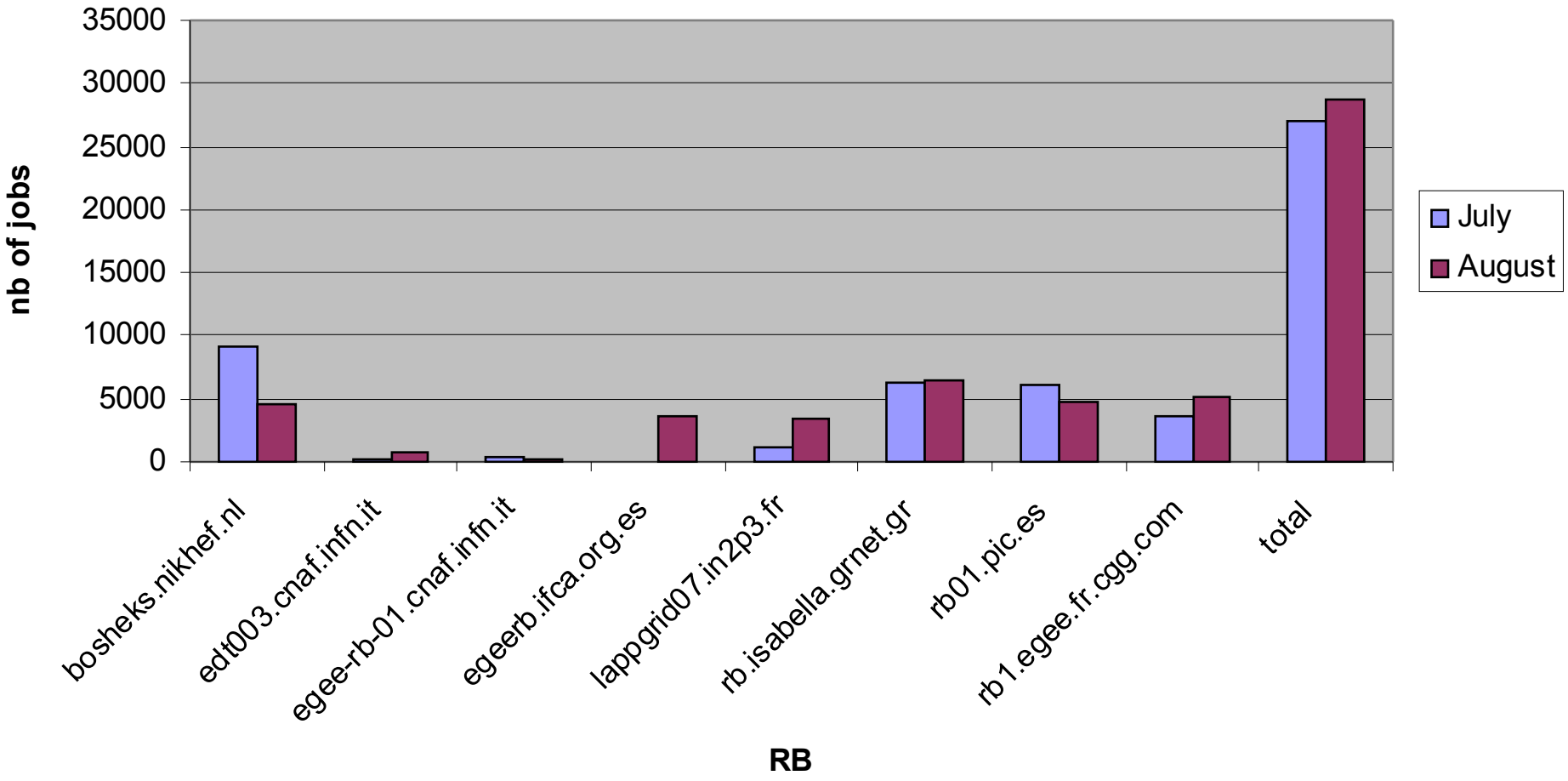
- **Biological information**
  - Plasmepsin is a promising aspartic protease target involved in the hemoglobin degradation of P. *falciparum*. 5 different structures are prepared (PDB source)
  - ZINC is an open source library of 3,3 millions selected compounds. They are made available by chemistry companies and are ready to be used
- **Biomedical informatics tools**
  - Autodock is free for academic, with grid based empirical potential and flexible docking via MC search and incremental construction
  - FlexX is a licensed software originally developed by Fraunhofer Institute, available for the data challenge during 2 weeks, with Boehm potential and fragment assembly energy function
- **Grid tools**
  - wisdom_env is an environment for an automatic, optimized and fault tolerance workflow using the grid resources and services
  - EGEE biomedical Virtual Organization will be dedicated/no-dedicated resources

# Molecular docking

- **Workload management**

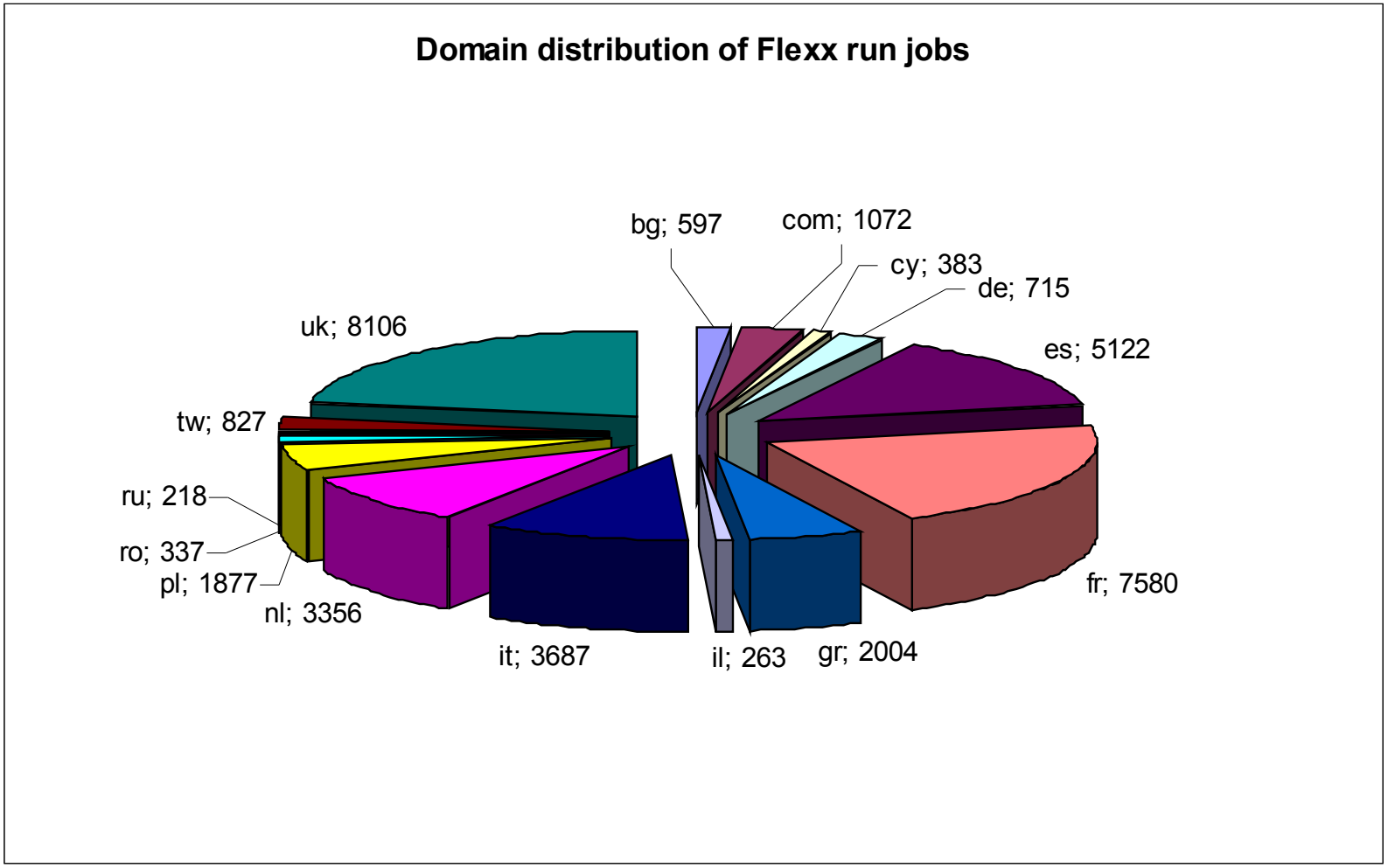


# Resources : JRA2 statistics

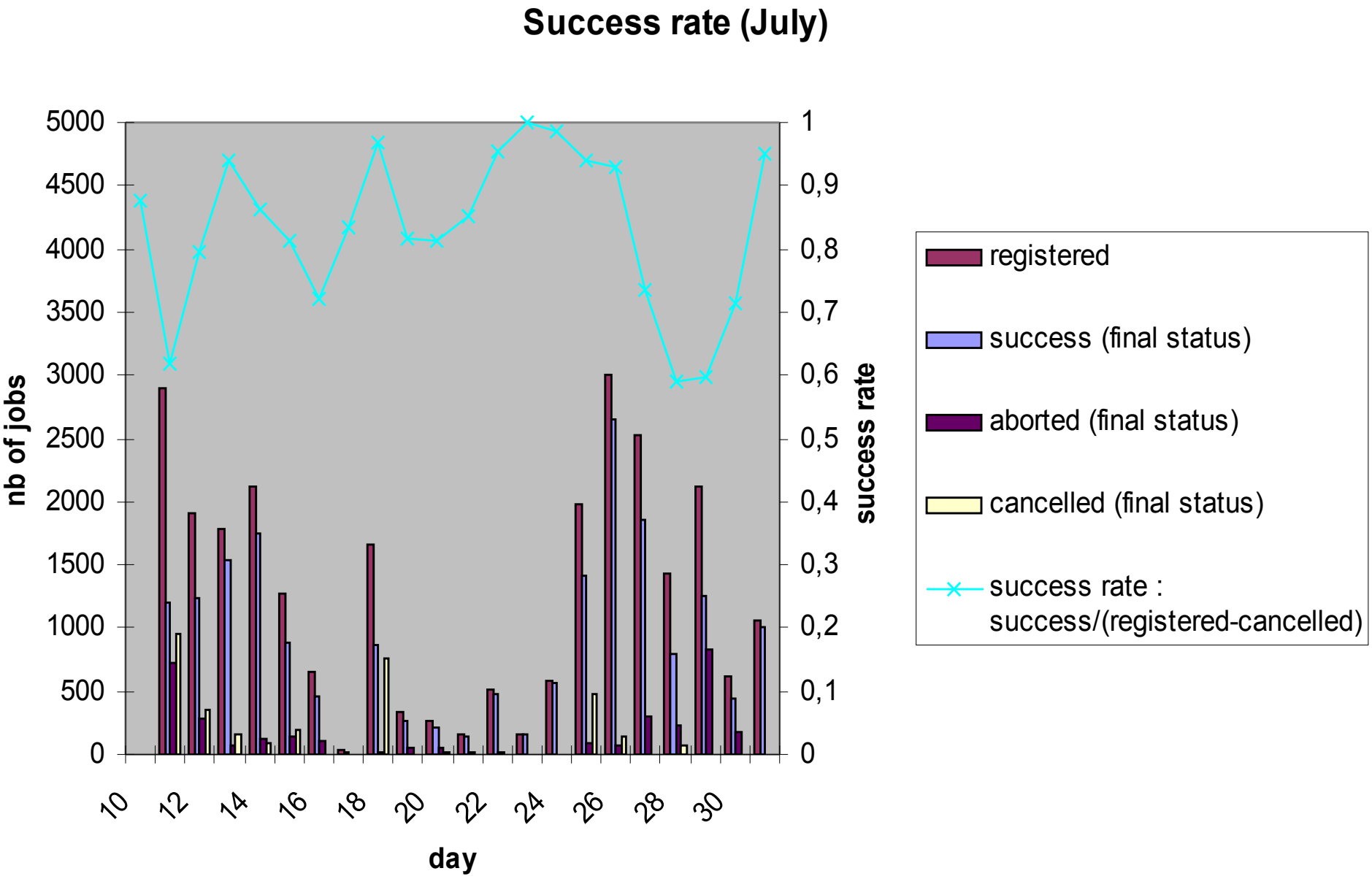


# Molecular docking (Data Challenge)

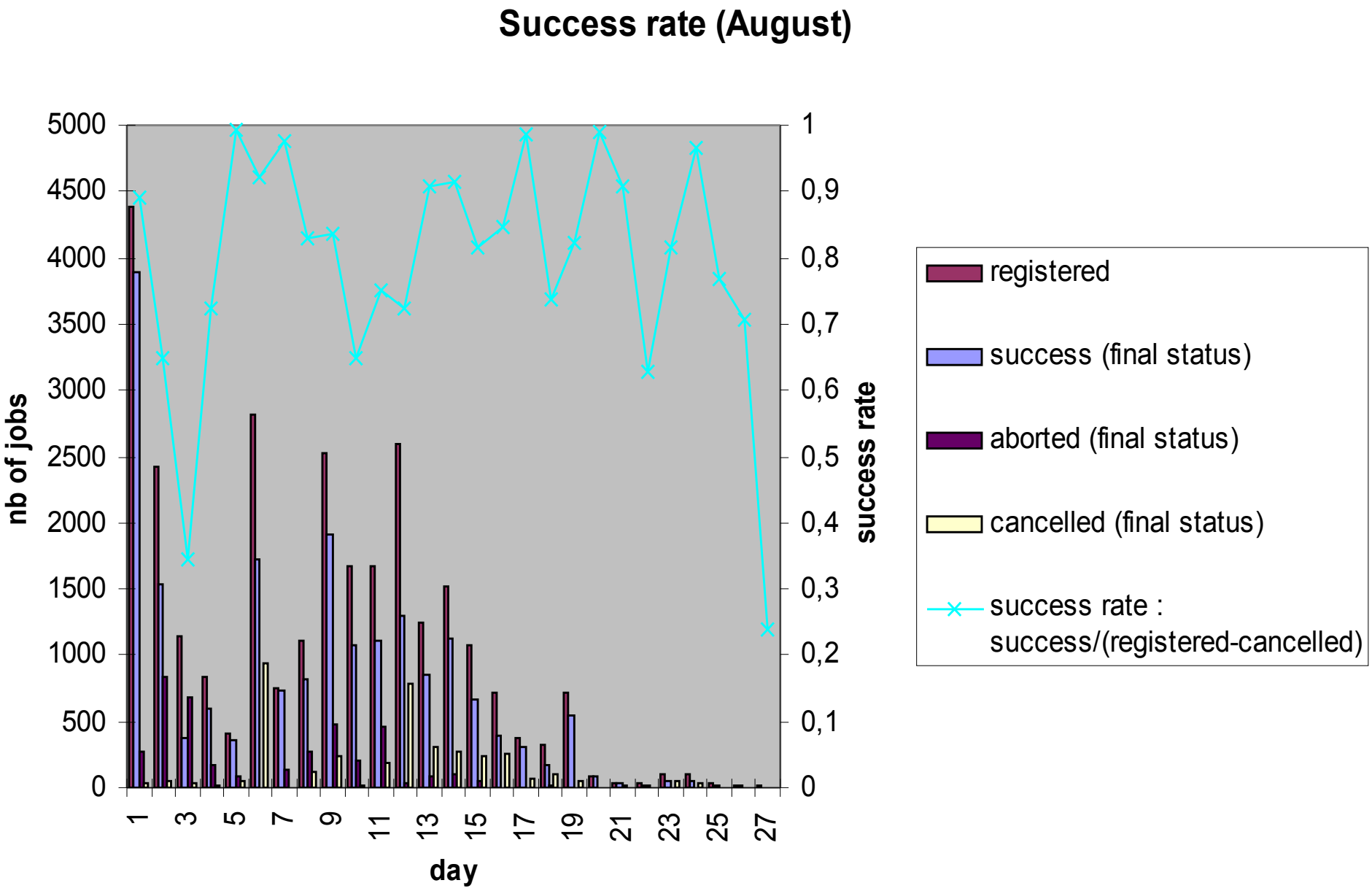
- International distribution of jobs



# Resources : JRA2 statistics

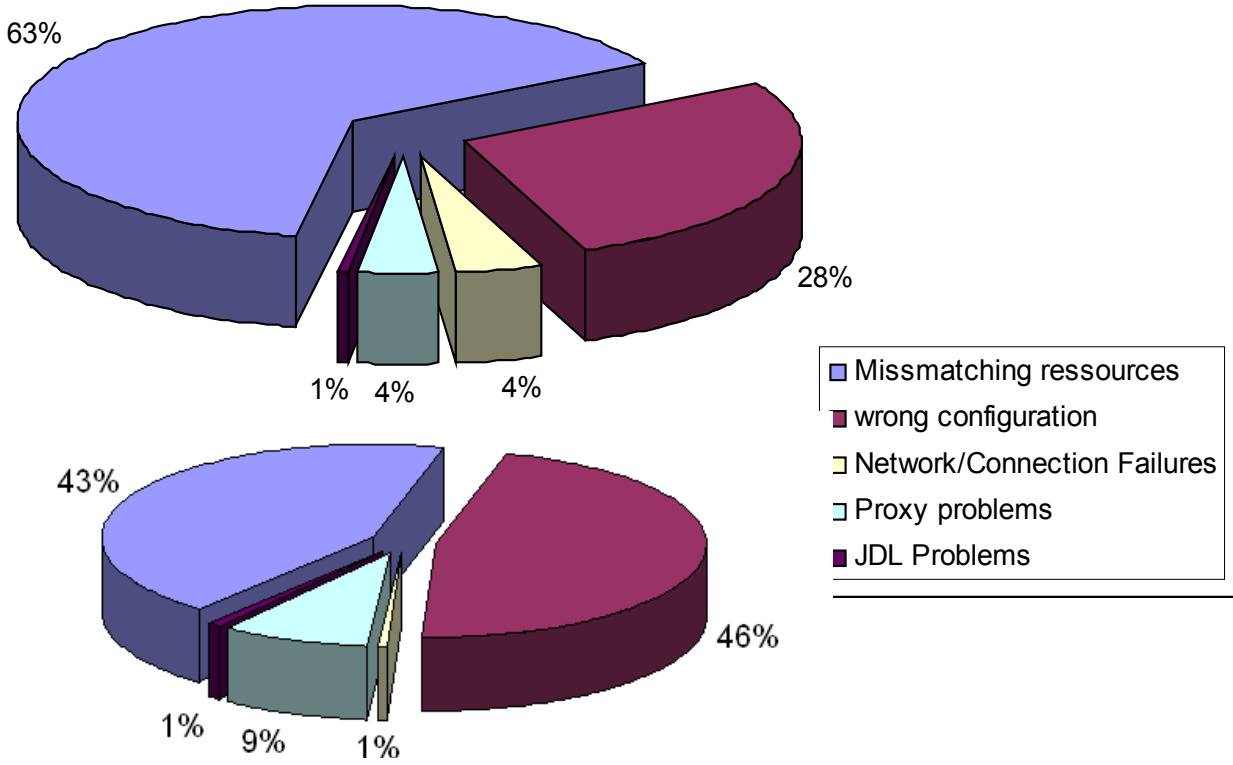


# Resources : JRA2 statistics



# Failures : JRA2 statistics

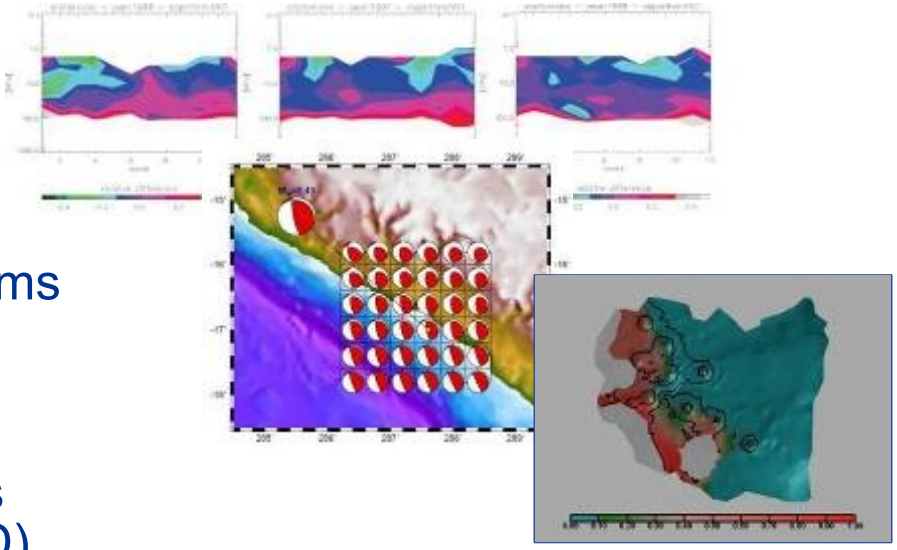
**Biomed data challenge**
**Abort reasons distribution**
**(10/07/2005 - 27-08-2005)**

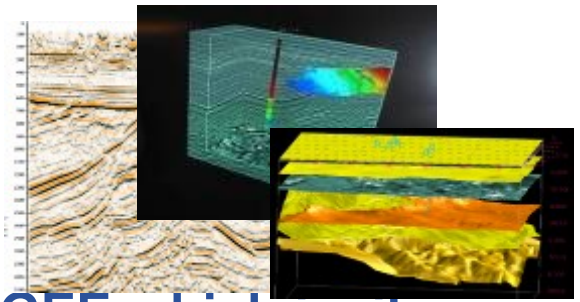



**Abort reasons distribution for all VO**
**01/2005 – 06/2005**

# 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

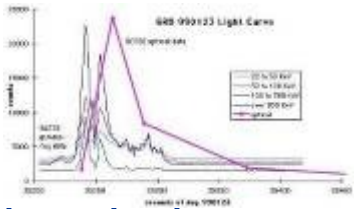


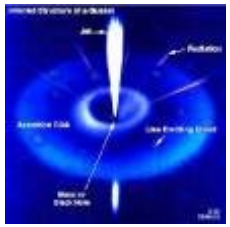


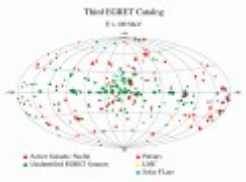
- **A large variety of applications ported on EGEE which incites new users**
- **Interactive Collaboration of the teams around a project**

# 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

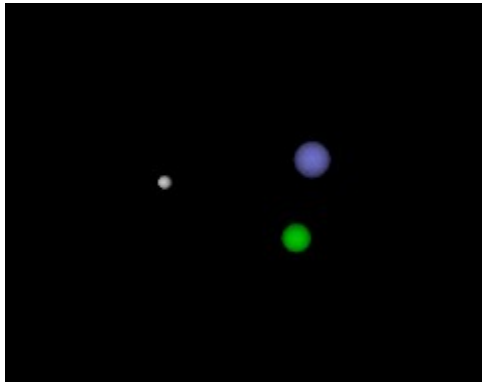


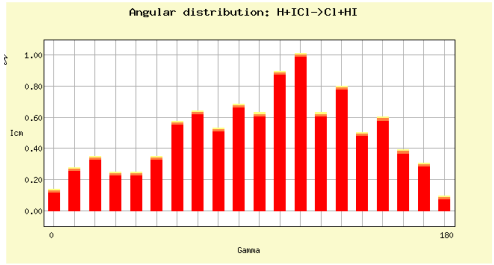


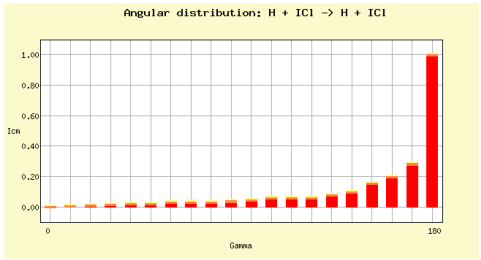


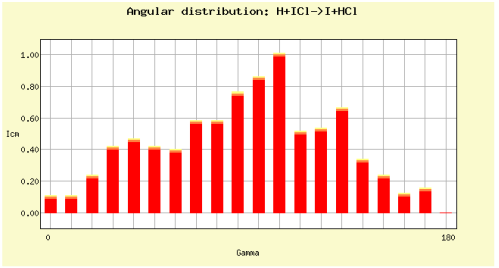
# Computational Chemistry

- **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



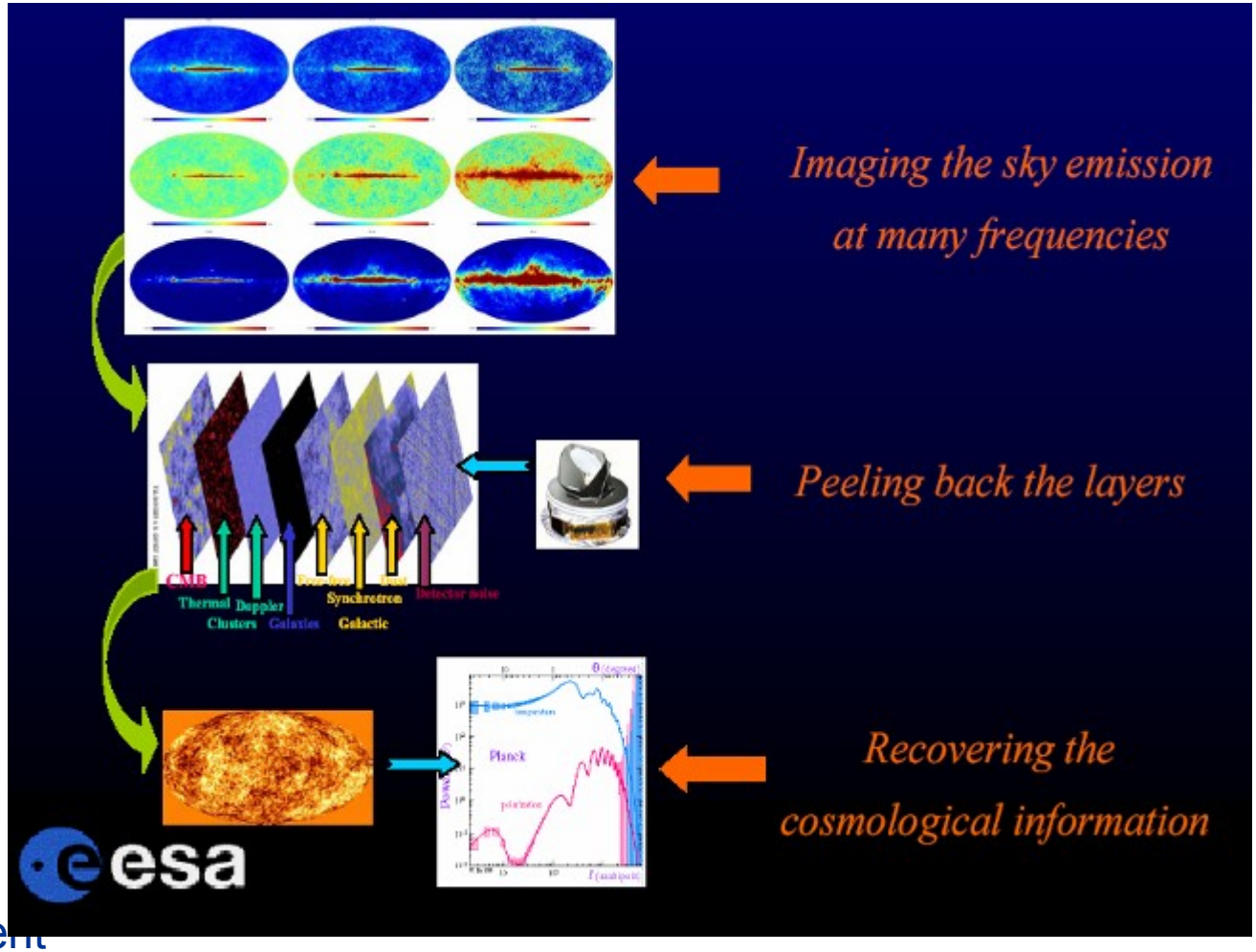


# Planck

- **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



# How to join ?

- **Different applications areas**
- **High Energy Physics: international consortium already joined!**
- **Health-related applications**
  - Biomed working group
    - Medical image processing
    - Bioinformatics
    - Molecular biology
  - Visit http://egee-na4.ct.infn.it/biomed/
- **Other applications**
  - “generic” application group
  - EGAAP (EGEE Generic Applications Advisory Panel) committee for selection and follow up
  - Training and induction courses
  - Visit http://egee-na4.ct.infn.it/generic/

# Important criteria

- **Scientific objectives**
- **User community**
- **Computing / data processing needs**
  - Are grids suitable?
  - What are the expected functionalities?
- **Application maturity and schedule**
- **Application support resources**
- **Resource provision**
- **There is no free lunch**
  - Use other's resources...
  - ...given that you provide some
- **Biomed and generic application questionnaires available from web sites.**

# Infrastructure future

- **EGEE: April 2004-March 2006**
- **EGEE2 proposal submitted (2006-2008)**
  - Increase in number of supported application areas
- **Longer term plans**
  - Future project taking over EGEE
  - Free license middleware
  - Large international adoption
  - Participation to standards
  - Established needs of application communities