Grids & E-Science

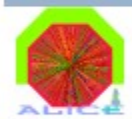
Fotis Georgatos <gef@grnet.gr>
Grid Technologies Trainer, GRNET

University of Patras, December 7th-8th, 2006

LCG/EGEE collaboration

Grid Projects Collaborating in LHC Computing Grid

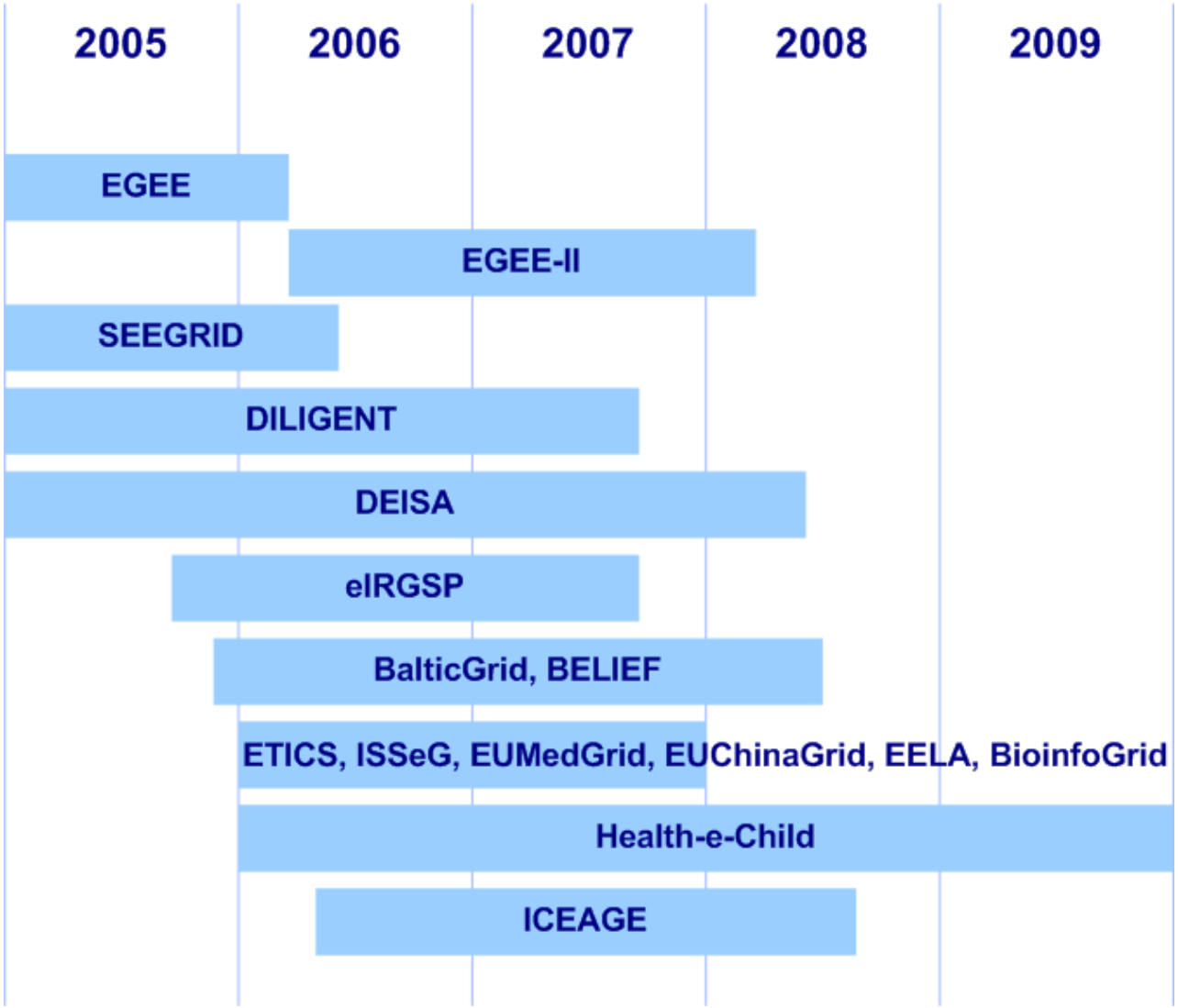
Open Science Grid



EGEE Operations Information	
Active Sites	~200
Available CPU	~30000
Available Storage (TB)	~10PBytes

European Grid projects, a timeline

Enabling Grids for E-sciencE



Computationally Intensive Science

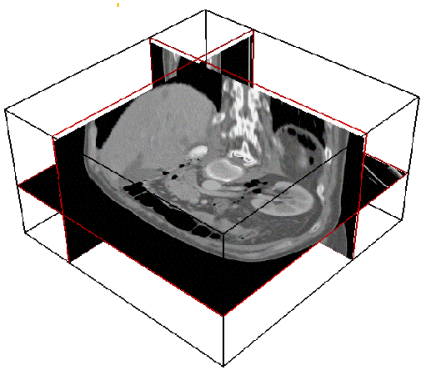
- **Science is becoming increasingly digital and needs to deal with increasing amounts of data**
- **Simulations get ever more detailed:**
  - Nanotechnology – design of new materials from the molecular scale
  - Modelling and predicting complex systems (weather forecasting, floods, earthquakes)
  - 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

Initial Grid cases: HEP & Biomed

High-Energy Physics (HEP)

- Requires 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)

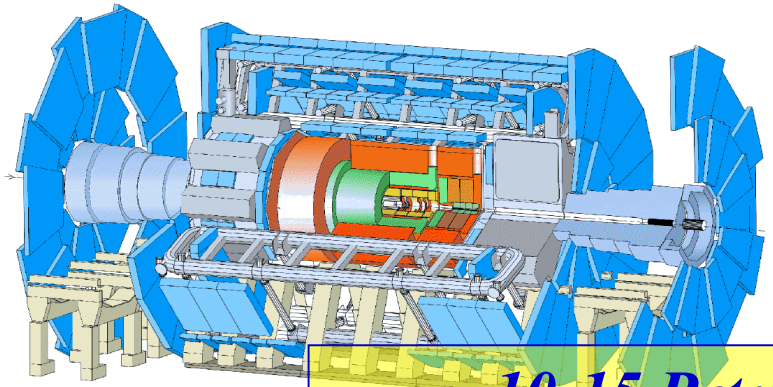
Biomedical Applications



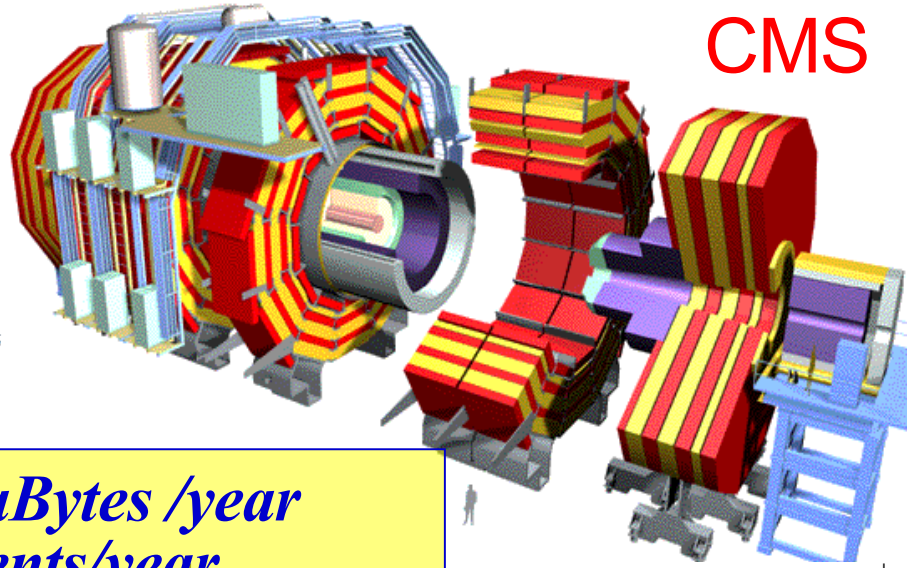
- Similar computing and data storage requirements
- Major additional challenge: security & privacy

HEP: LHC collaborations

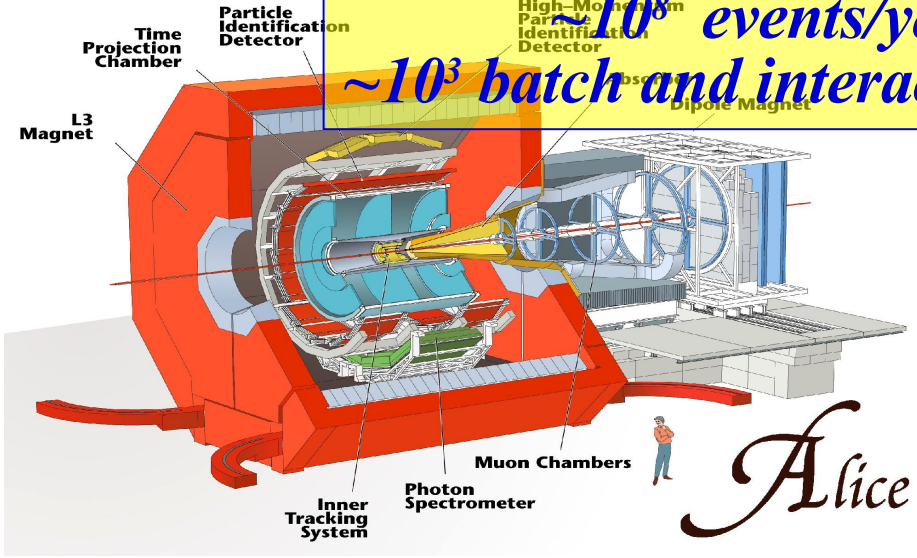
ATLAS

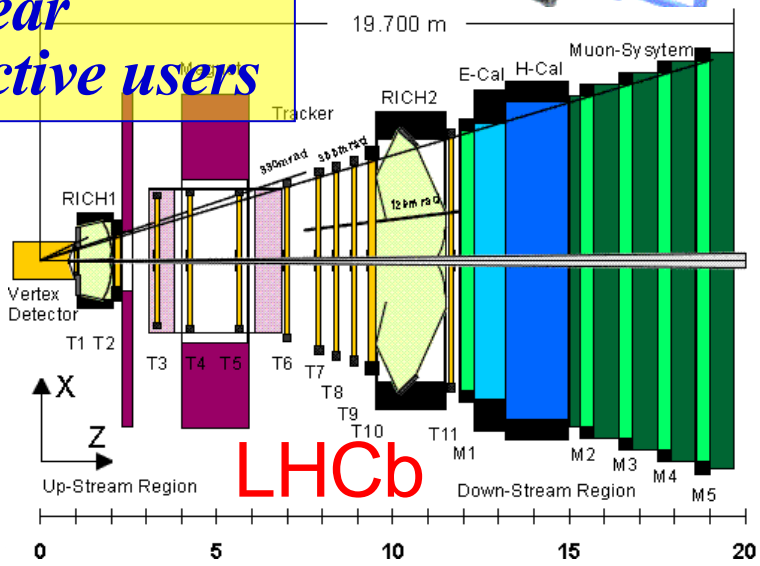


CMS



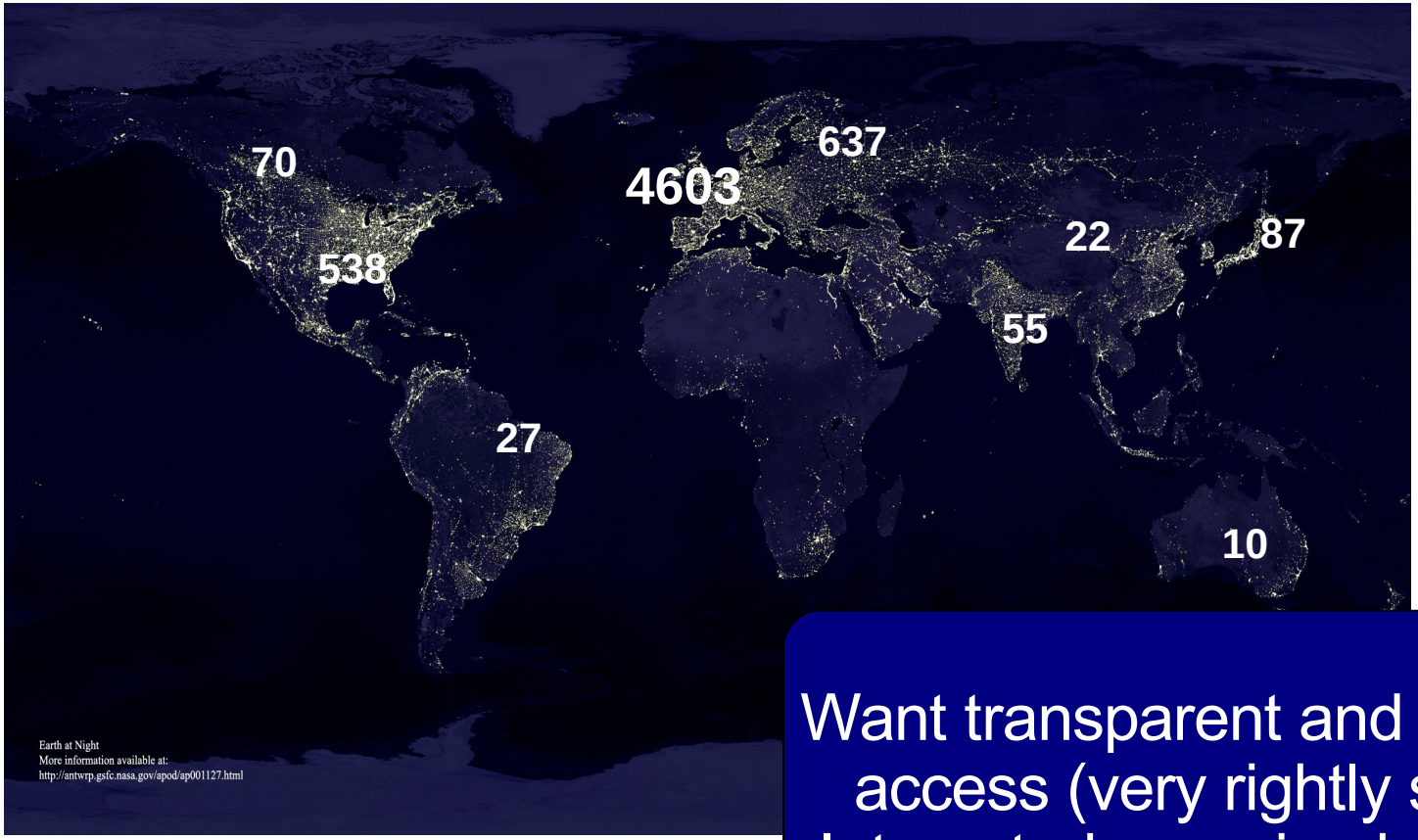
~10-15 PetaBytes /year
~10^8 events/year
~10^3 batch and interactive users





LHC: Big "Big Science"

Over 6000 LHC Scientists world wide

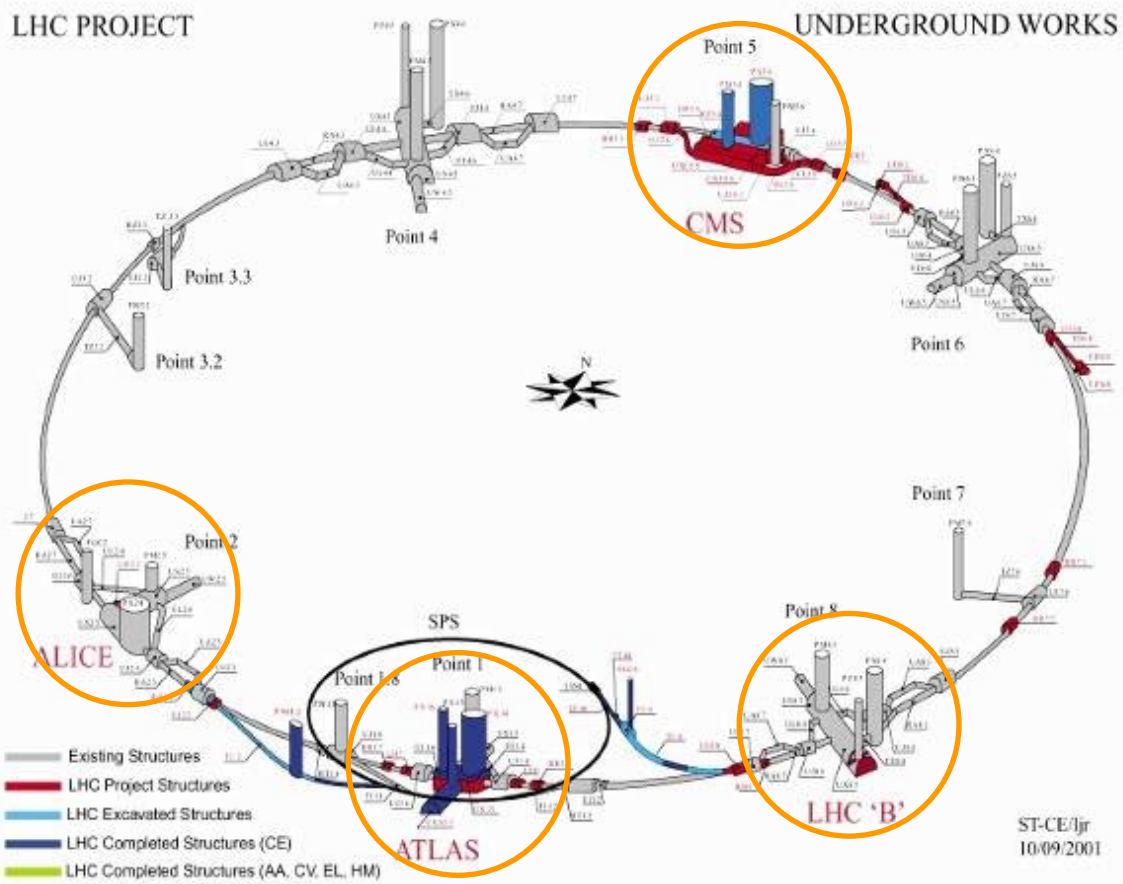


Europe: 267 Institutes, 4603 Users
Other: 208 Institutes, 1632 Users

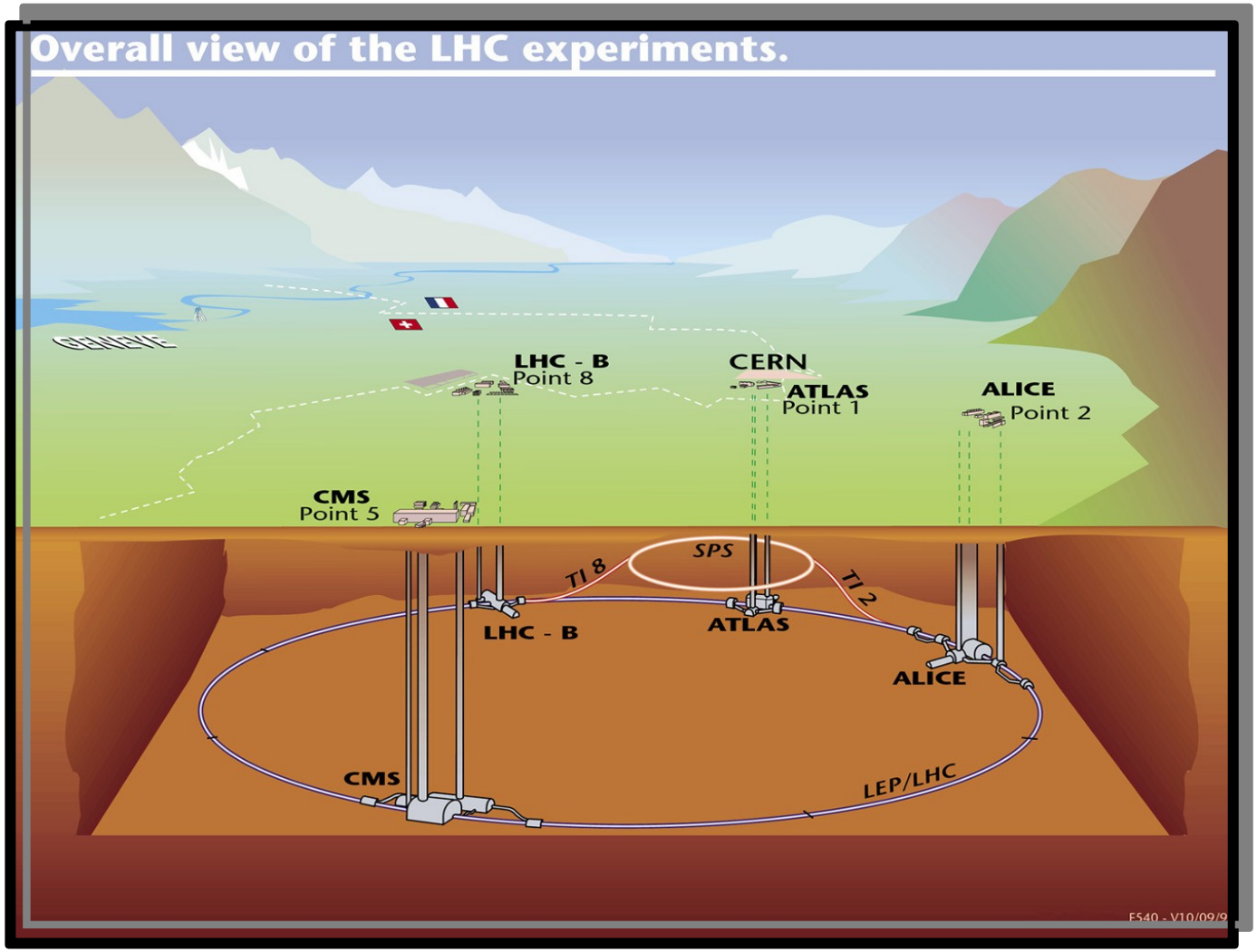
Want transparent and quick access (very rightly so). Interested more in physics results, than computing revolutions

The LHC experiment

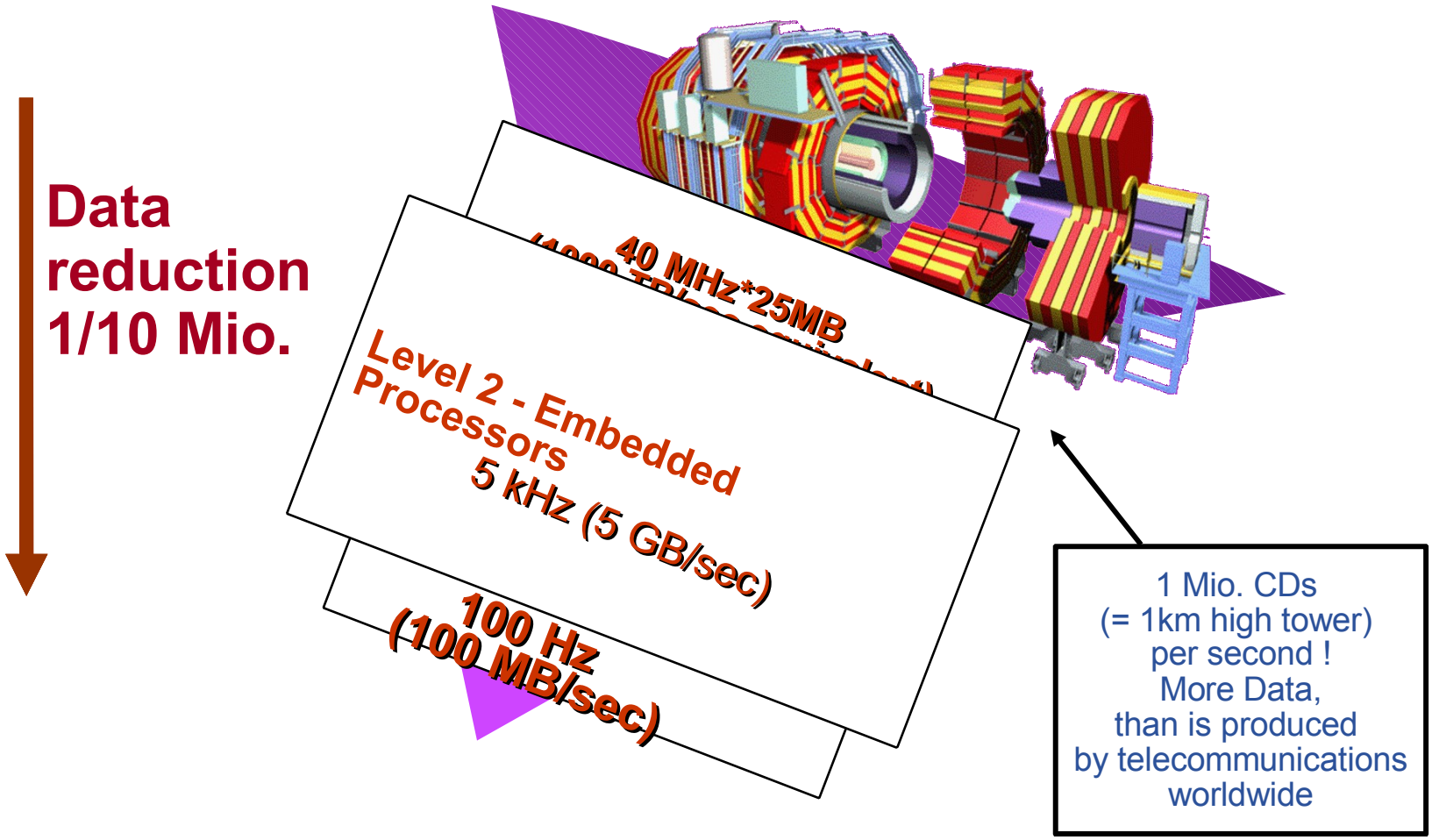
- **Large Hadron Collider**
  - Four experiments:
    - ALICE
    - ATLAS
    - CMS
    - LHCb
  - 27 km tunnel
  - Start-up in 2007



The underground LHC facility



LHC, the CMS detector

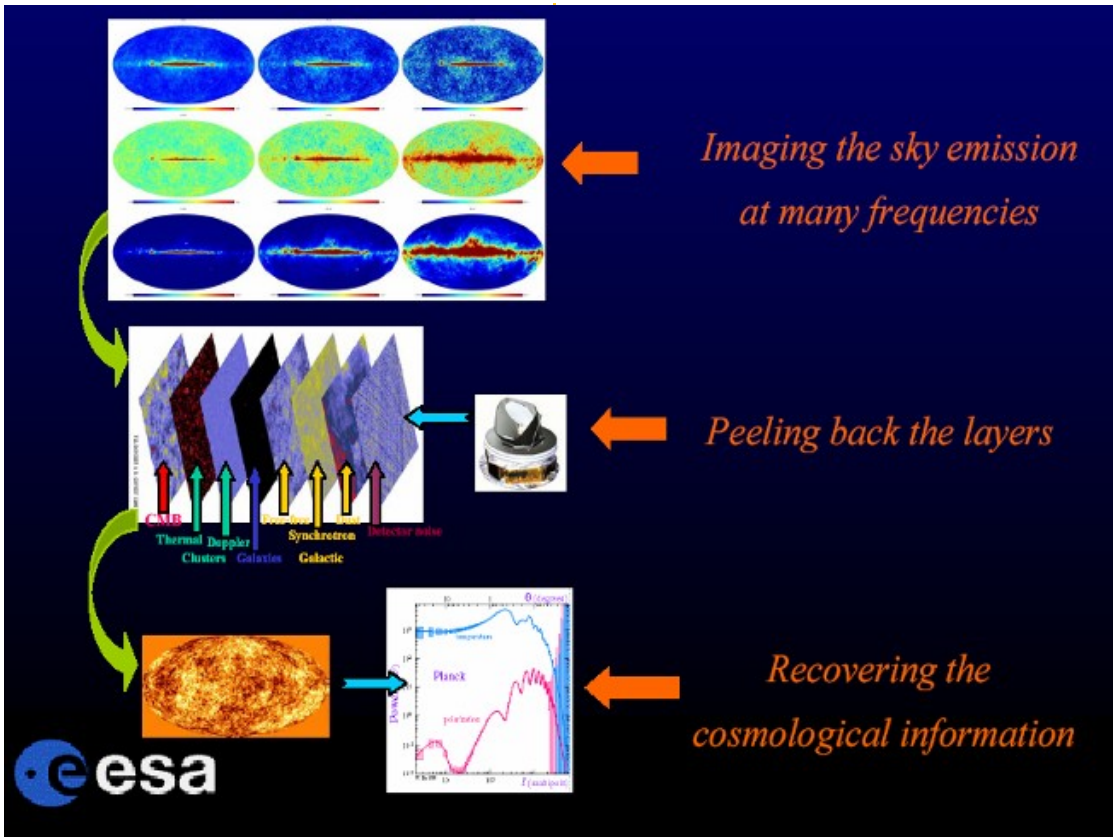


1 PB of data per year and experiment

... and 6000 physicist that want to access it !

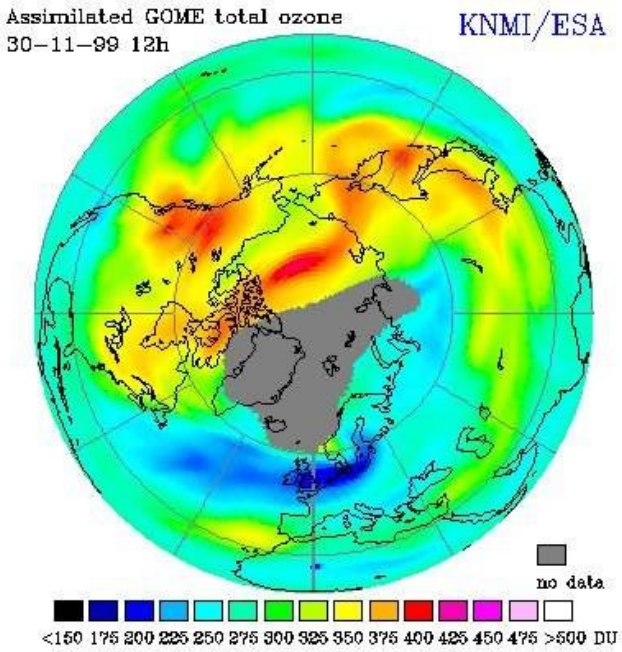
HEP case: Planck experiment

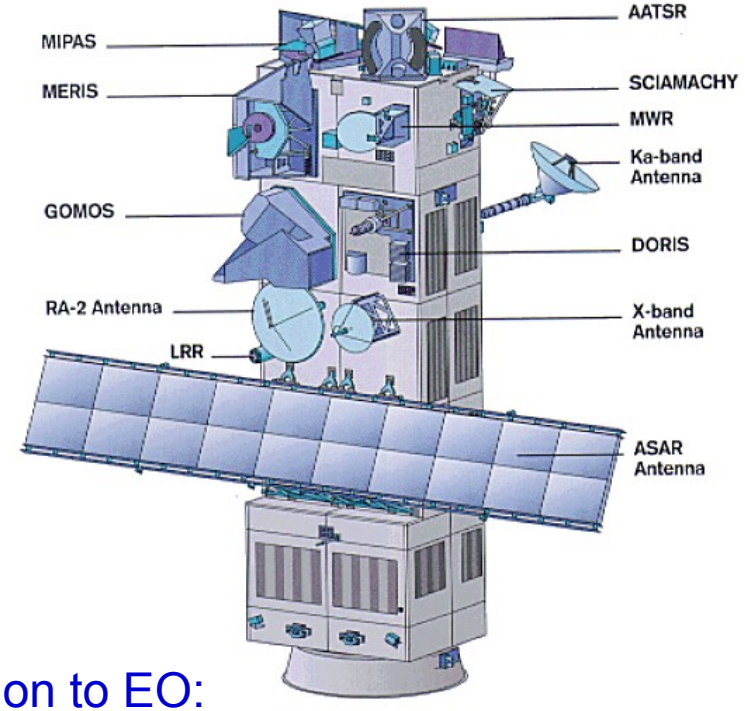
- **On the Grid:**
  > 12 time faster
  (only ~5% failures)
- **Complex data structure**
  → data handling important
- **The Grid as**
  - Collaboration tool
  - common user-interface
  - flexible environment
  - new approach to data and S/W sharing



Earth Observation: success cases

ESA missions:
100's of Gbytes of data per day



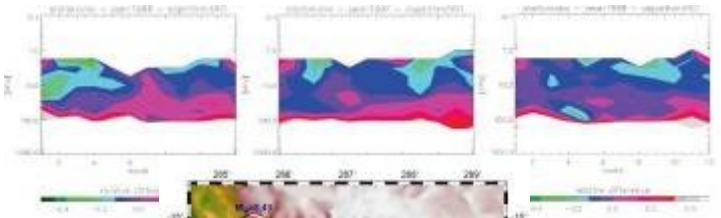


Grid contribution to EO:
Enhance the ability to access high level products
Allow reprocessing of large historical archives
Improve Earth science complex applications (data fusion, data mining, modelling …)

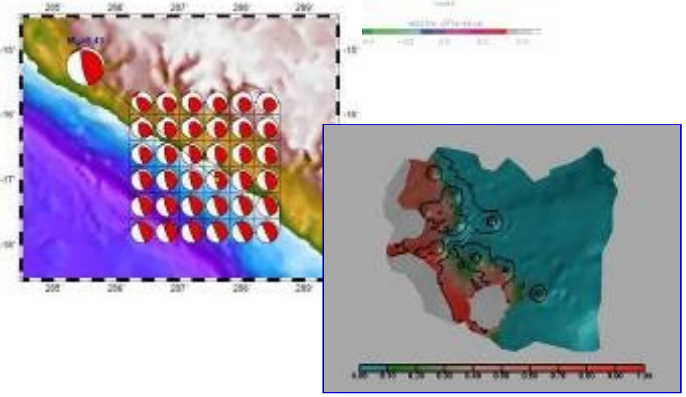
Federico.Carminati , EU review presentation, 1 March 2002

Earth Sciences

- **Earth Observations by Satellite**
  - Ozone profiles

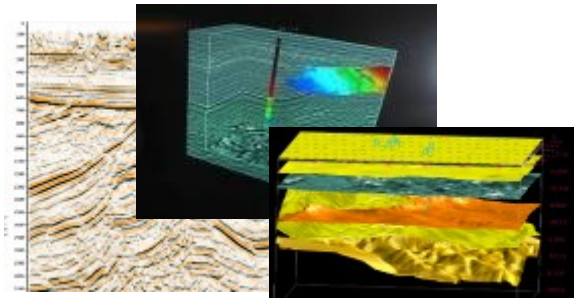


- **Solid Earth Physics**
  - Fast Determination of mechanisms of important earthquakes



- **Hydrology**
  - Management of water resources in Mediterranean area (SWIMED)

- **Geology**
  - Geocluster: R&D initiative of the Compagnie Générale de Géophysique

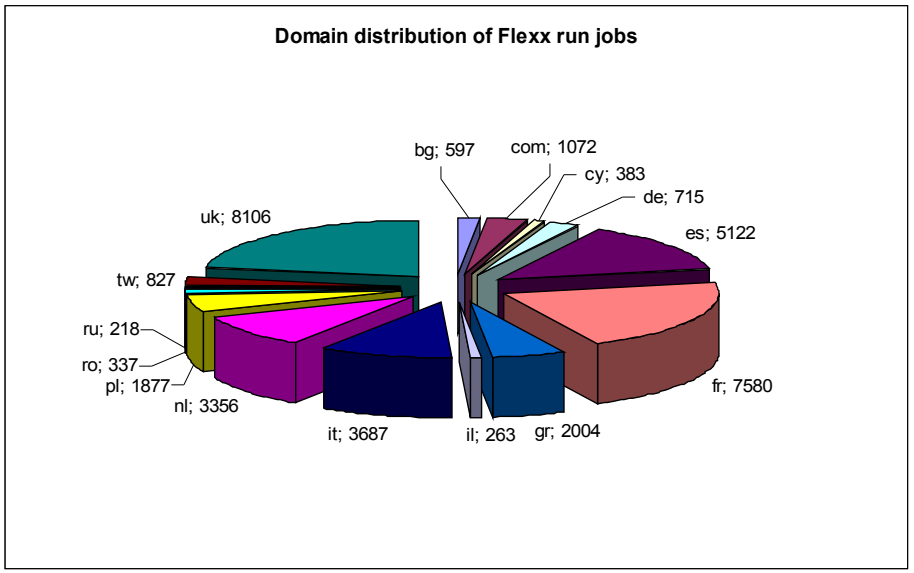


- **A large variety of applications ported on EGEE which incites new users**
- **Interactive Collaboration of the teams around a project**

Biomed case: WISDOM

World-wide In-silico Docking on Malaria

- Significant biological parameters
  - two different molecular docking applications (Autodock & 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
- **Next case: SARS, H5N1 research on the grid!**



WISDOM open day

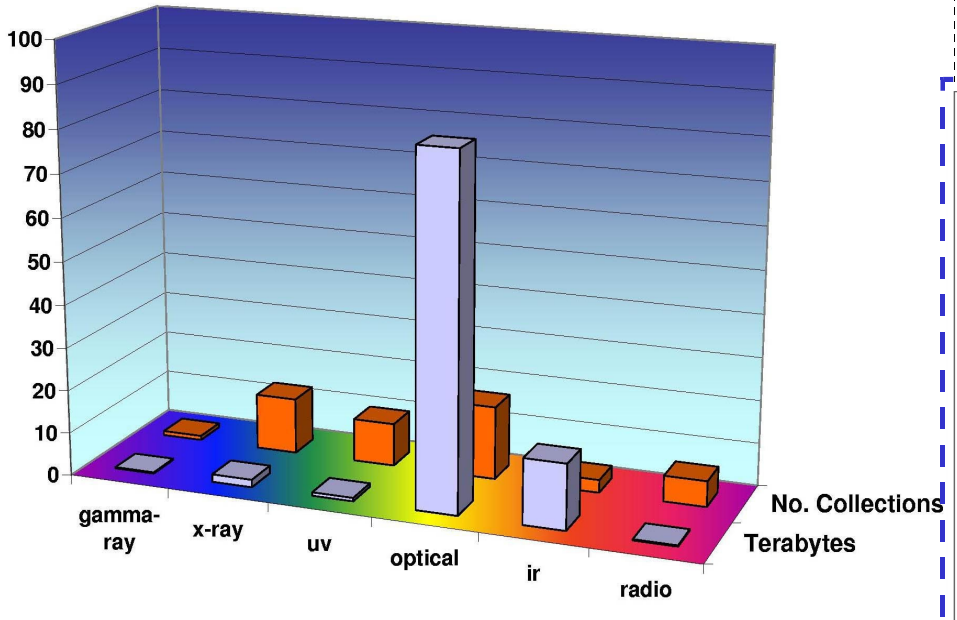
December 16th, 2005, Bonn (Germany)

Discuss Data Challenge results

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

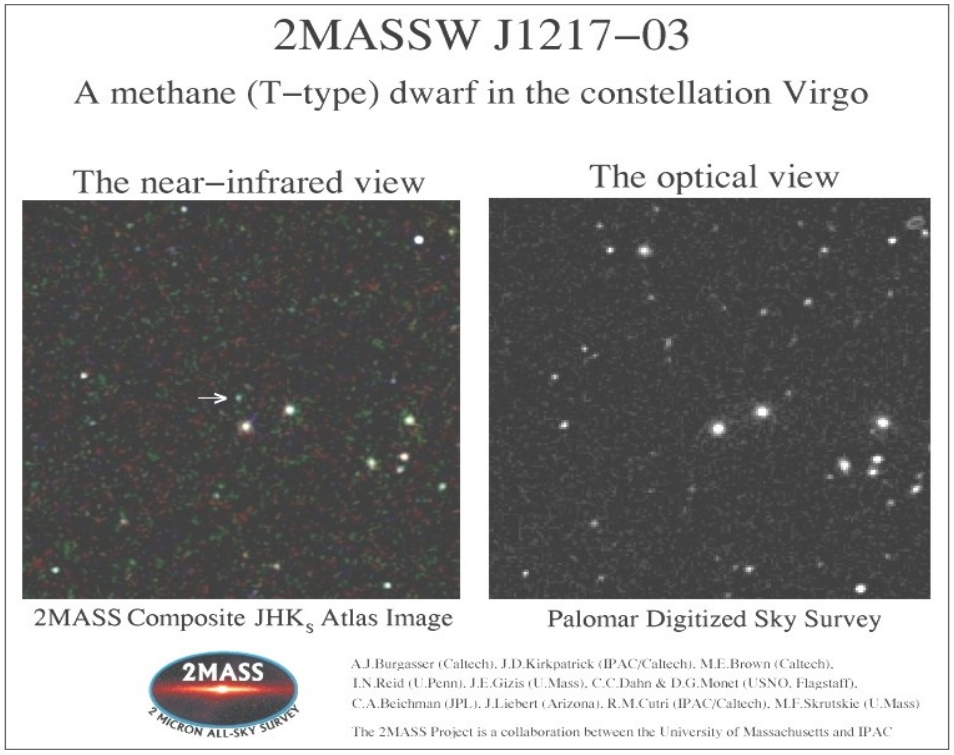
Information: http://wisdom.eu-egee.fr

Astronomy



No. & sizes of data sets as of mid-2002, grouped by wavelength

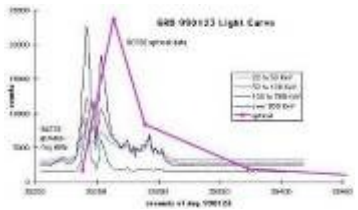
- 12 waveband coverage of large areas of the sky
- Total about 200 TB data
- Doubling every 12 months
- Largest catalogues near 1B objects

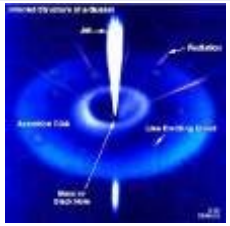


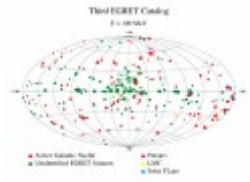
Data and images courtesy Alex Szalay, John Hopkins University

MAGIC telescope and VO

- **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





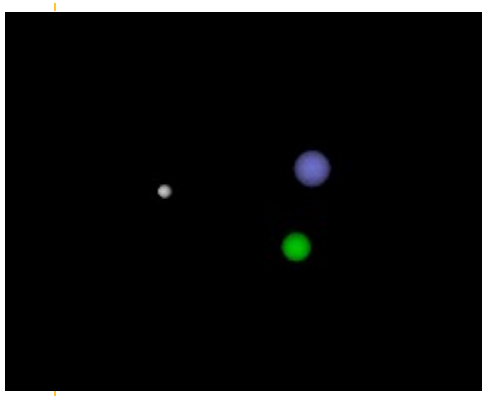


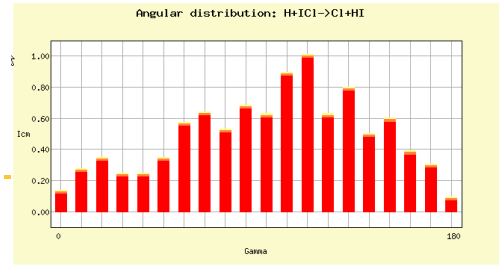
Computational Chemistry

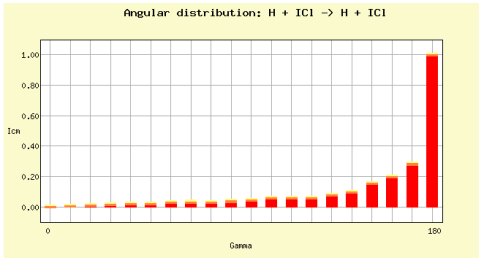
The Grid Enabled Molecular Simulator (GEMS)

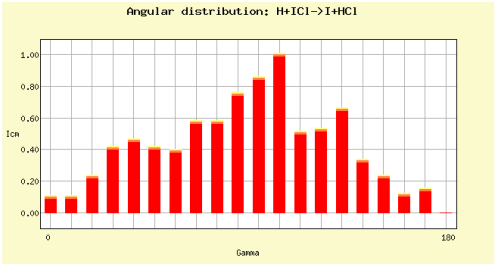
- Motivation:
  - Modern computer simulations of biomolecular systems produce an abundance of data, which could be reused several times by different researchers.
    → data must be catalogued and searchable

- GEMS database and toolkit:
  - autonomous storage resources
  - metadata specification
  - automatic storage allocation and replication policies
  - interface for distributed computation



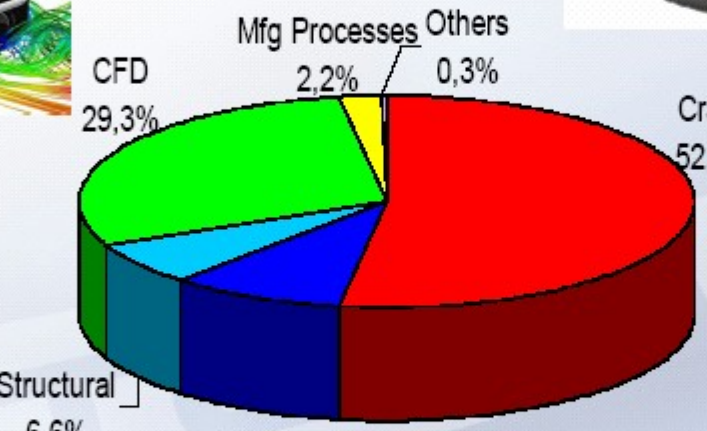


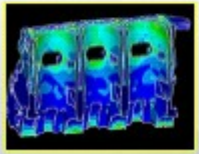


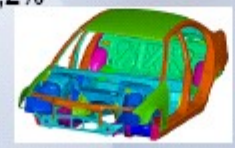


Automotive Industry: Grid potential

HPC Application Segments in Automotive



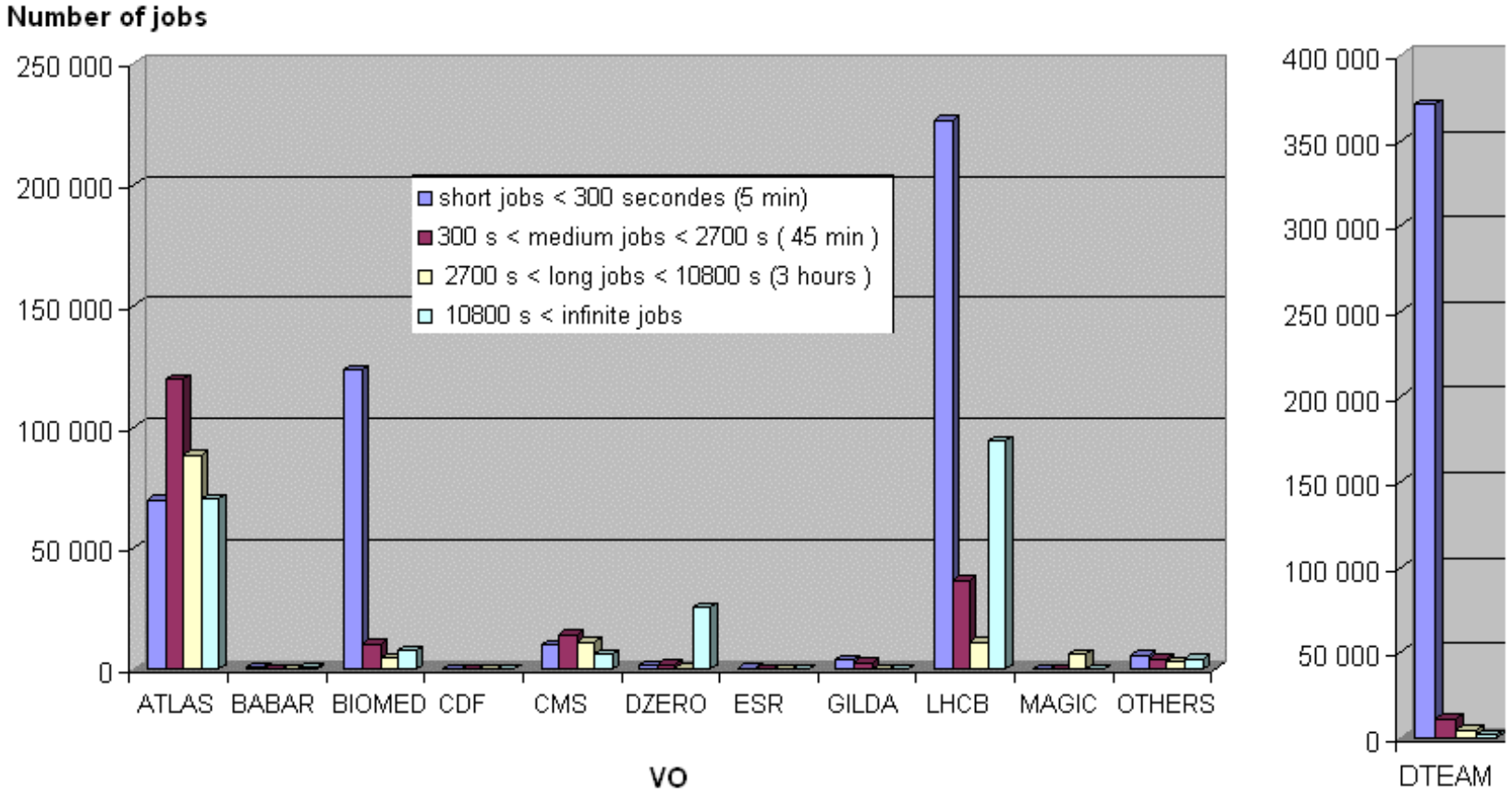




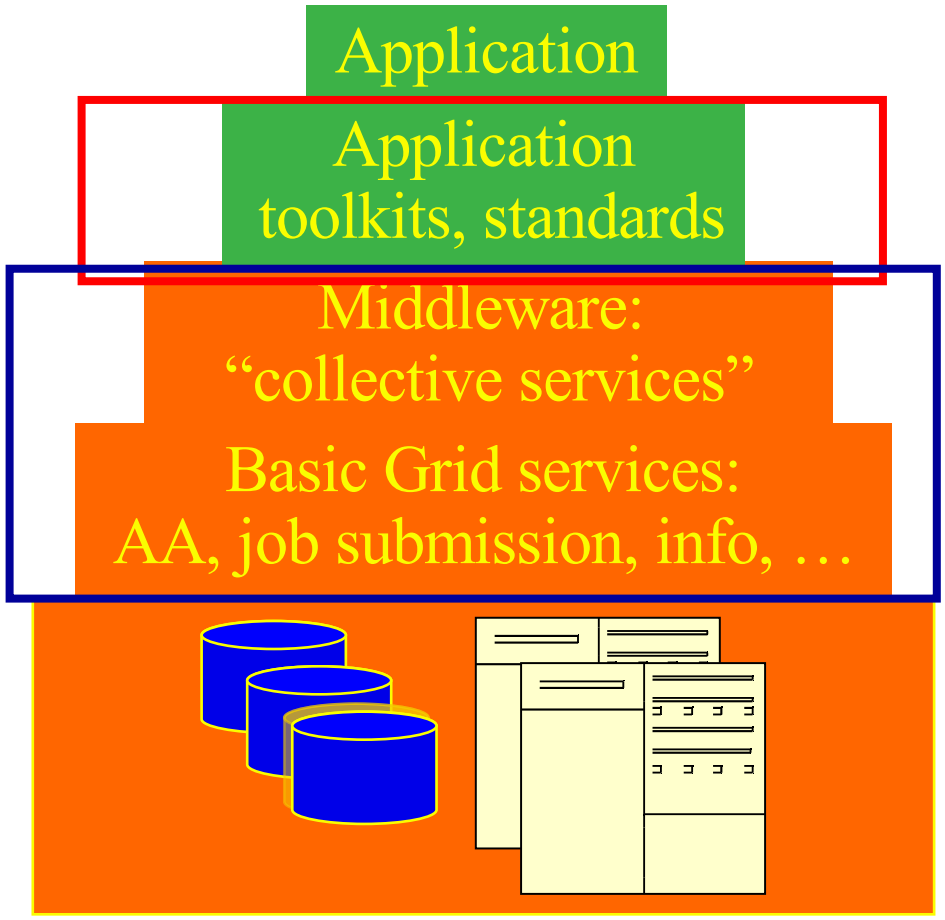
- Crash is the application segment #1
- CFD is the fastest growing application segments (5% yty])
- NVH the most demanding in terms of memory and IO bandwidth.

EGEE infrastructure usage

Average job duration
January 2005 – June 2005 for 10 major VOs



Future: Empowering VOs



VO-specific developments:

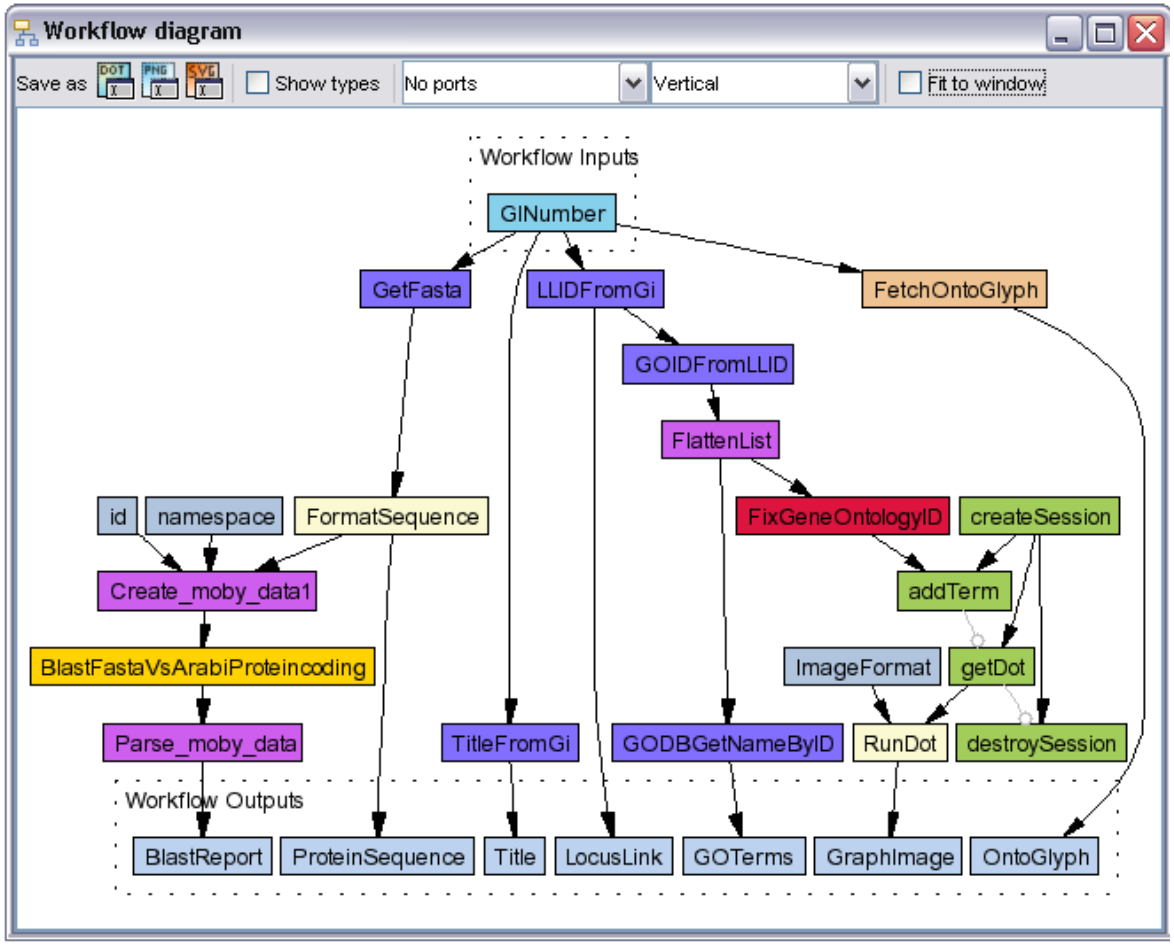
- **Portals**
- **Virtual Research Environments**
- **Semantics, ontologies**
- **Workflow**
- **Registries of VO services**

Production grids provide these services.

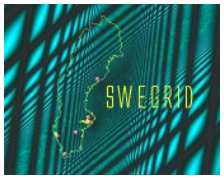
Develop above these to empower non-UNIX specialists!

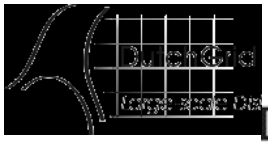
Future: Workflow features

- Taverna in MyGrid http://www.mygrid.org.uk/
- "allows the e-Scientist to describe and enact their experimental processes in a structured, repeatable and verifiable way"
- GUI
- Workflow language
- Enactment engine



National Grid initiatives





CroGrid

E-sciencE conclusions...

- We live in a time where the computing infrastructure makes **distributed computation more attractive** than centralised computation – at least for some applications
- Many scientific disciplines, application areas and organisation types create a **demand for a global computing infrastructure**
- **Grid Computing has gained a lot of momentum**, its meaning has started to change
- As explained, this gain in momentum stems from the drastically **increased hardware capabilities and new application types**
- The **theoretical groundwork** for a distributed computing infrastructure has been available since long time – distributed computing and Grid computing is not really a new phenomenon (**only the name is new, plus a couple of facilities**)
- **The challenge in building this infrastructure lies in the large scale and in the need for standardisation and bridge building**


If “The Grid” vision leads us here…
… then where are we now?

Q & A