EGEE: An e-Infrastructure for Science

Frank Harris

CERN/Oxford

Ju

www.eu-egee.org

Information Society and Media

Overview of talk

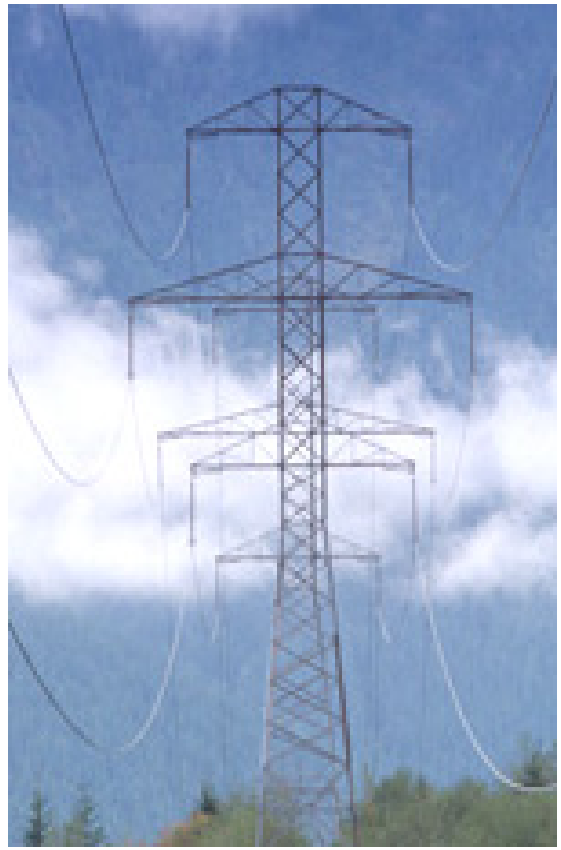
- **A brief overview of ‘what is the Grid’**
- **Aims of the EGEE project**
  - the services provided by the infrastructure
- **Current status**
  - emphasis on use by wide range applications
- **Summary comments on EGEE, other grid projects and ‘the future’**
- **How can you learn more and get hands-on experience?**
- **Questions…….**

What is the Grid?

- The World Wide Web provides seamless access to information that is stored in many millions of different geographical locations
- In contrast, the Grid is a new computing infrastructure which provides seamless access to computing power and data distributed over the globe
- The name Grid is chosen by analogy with the electric power grid: plug-in to computing power without worrying where it comes from, like a toaster

  (Foster and Kesselman 1997 – grid pioneers- wrote seminal book in 1998- and many papers
  - http://www.globus.org/alliance/publications/papers/anatomy.pdf

  'Anatomy of the grid - enabling scalable virtual organisations', 2001)



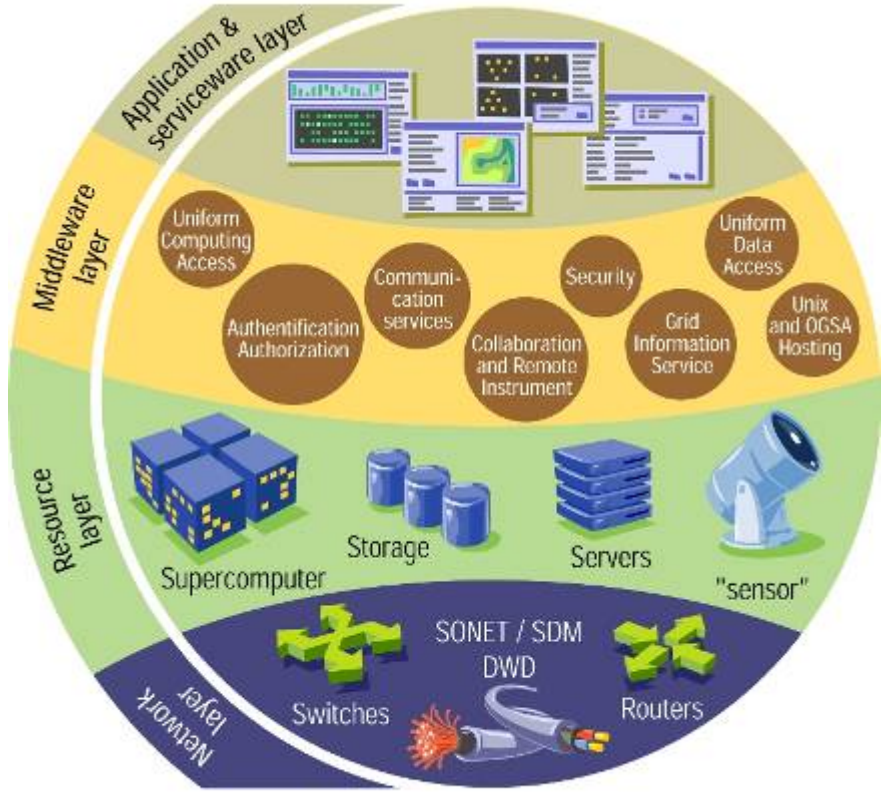
- **The Grid allows (quote from Anatomy of grids paper)**

‘coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organisations’

Resources: computers, data, software, collaborative tools…

How does the Grid work?

- **It relies on advanced software, called** middleware**.**
- **Middleware automatically finds the** data **the scientist needs, and the** computing power **to analyse it.**
- **Middleware balances the load on different resources. It also handles** security, accounting, monitoring **and much more.**



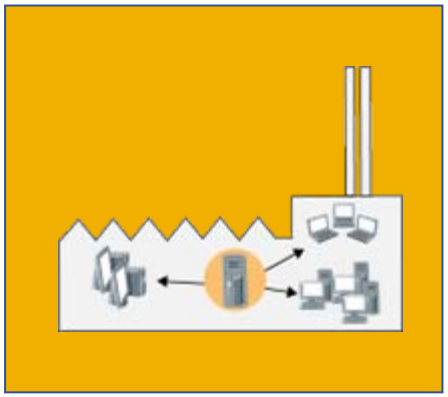
Virtualisation & Sharing

- **Virtual Organisations**
  - People from different organisations but with common goals get together to solve their problems in a cooperative way – e.g an HEP experiment or a Biomedical organisation
- **Virtualised shared computing resources**
  - Members of VOs have access to computing resources outside their home institutions. Resource providers typically have a contract/MoU with the VO, not with the VO members
- **Virtualised shared data resources**
  - Similar to computing resources
- **Other resources may be shared and virtualised as well:**
  - Instruments, sensors, even people

Virtualization of resources is needed to abstract from their heterogeneity

Different Grids for different needs

- **There is as yet no unified Grid (like there is a single web) rather** there are many Grids **for many applications.**
- **The word Grid is used to signify different types of distributed computing for example** Enterprise Grids **(within one company) and** public resource Grids **(volunteer your own PC).**
- **In this talk, focus is on** scientific Grids **that link together major computing centres in research labs and universities.**
- **Latest trend is to** federate national Grids **to achieve a global Grid infrastructure. High Energy Physics is a driving force for this.**



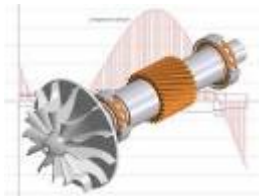
Grids serving science

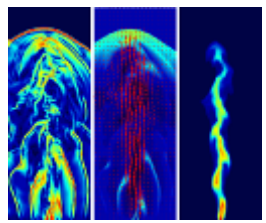
- Physics/Astronomy ***(data from different kinds of research instruments)***
- Medical/Healthcare ***(imaging, diagnosis and treatment)***
- Bioinformatics ***(study of the human genome and proteome to understand genetic diseases)***
- Nanotechnology ***(design of new materials from the molecular scale)***
- Engineering ***(design optimisation, simulation, failure analysis and remote Instrument access and control)***
- Natural Resources and the Environment ***(weather forecasting, earth observation, modeling and prediction of complex systems: river floods and earthquake simulation)***

The EGEE project

(preceded by Datagrid project Jan 2001-Mar 2004)

- **EGEE**
  - 1 April 2004 – 31 March 2006
  - 71 partners in 27 countries, federated
- **EGEE-II**
  - 1 April 2006 – 31 March 2008
  - 91 partners in 32 countries
  - 13 Federations
- **Objectives**
  - Large-scale, production-quality infrastructure for e-Science
  - Attracting new resources and users from industry as well as science
  - Improving and maintaining “gLite” Grid middleware





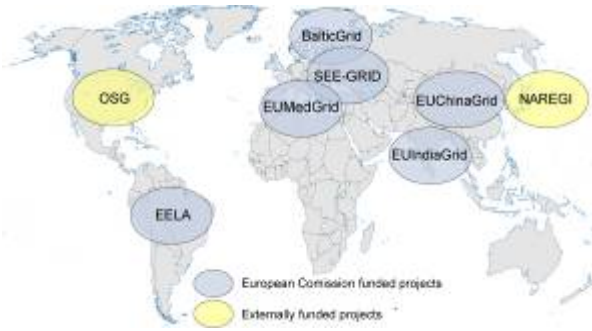
EGEE – The services provided to users

- **Infrastructure operation**
  - Currently includes 180+ sites across 40 countries
  - Continuous monitoring of grid services & automated site configuration/management
    http://gridportal.hep.ph.ic.ac.uk/rtm/launch_frame.html

- **Middleware**
  - Production quality middleware distributed under business friendly open source licence

- **User Support**
  - Training
  - Expertise in grid-enabling applications
  - Online helpdesk
  - Networking events (User Forum, Conferences etc.)

- **Interoperability**
  - Expanding geographical reach and interoperability with related infrastructures



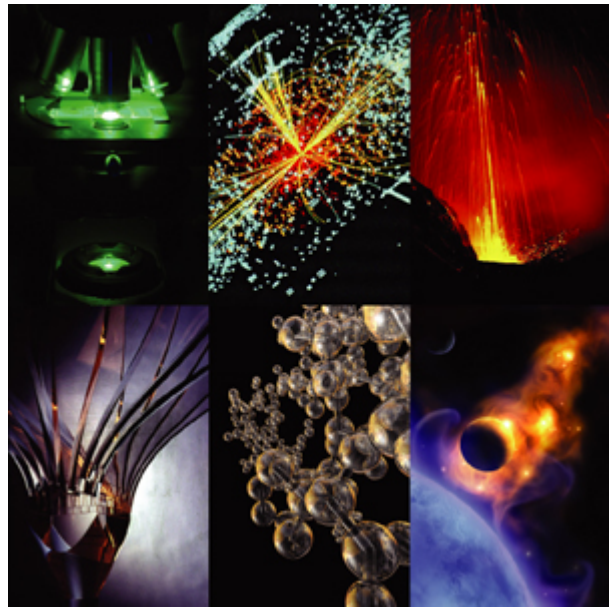
EGEE:
Steady growth over the lifetime of the project

EGEE:
> 180 sites, 40 countries
> 24,000 processors,
~ 5 PB storage

country	sites	country	sites	country	sites
Austria	2	India	2	Russia	12
Belgium	3	Ireland	15	Serbia	1
Bulgaria	4	Israel	3	Singapore	1
Canada	7	Italy	25	Slovakia	4
China	3	Japan	1	Slovenia	1
Croatia	1	Korea	1	Spain	13
Cyprus	1	Netherlands	3	Sweden	4
Czech Republic	2	FYROM	1	Switzerland	1
Denmark	1	Pakistan	2	Taipei	4
France	8	Poland	5	Turkey	1
Germany	10	Portugal	1	UK	22
Greece	6	Puerto Rico	1	USA	4
Hungary	1	Romania	1	CERN	1

Why do people work with EGEE?

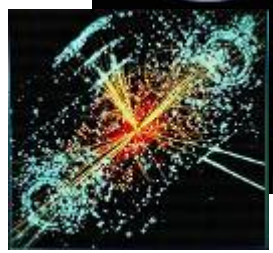
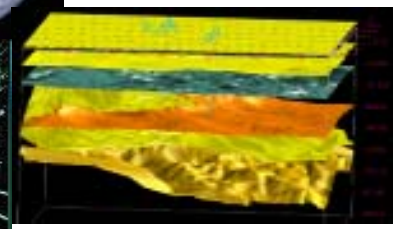
- **A global Grid infrastructure helps to provide easier access to resources for**
  - Small research groups
  - Scientists from many different fields
  - Remote and still developing countries
- **Users have access to new technologies**
  - Produce and store massive amounts of data
  - Transparent access to millions of files across different administrative domains
  - Low cost access to large computing resources
    - Mobilise large amounts of CPU & storage on short notice
  - High-end facilities
- **And users find new ways to collaborate**
  - Develop applications using distributed complex workflows
  - Eases distributed collaborations
  - New modes of community building
  - Easier access to higher education

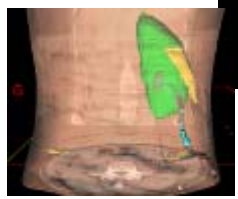


Current status of applications on EGEE

- **Applications from an increasing number of domains (there are ~70 VOs in EGEE)**
  - Astrophysics
  - Computational Chemistry
  - Earth Sciences
  - Financial Simulation
  - Fusion
  - Geophysics
  - High Energy Physics
  - Life Sciences
  - Multimedia
  - Material Sciences



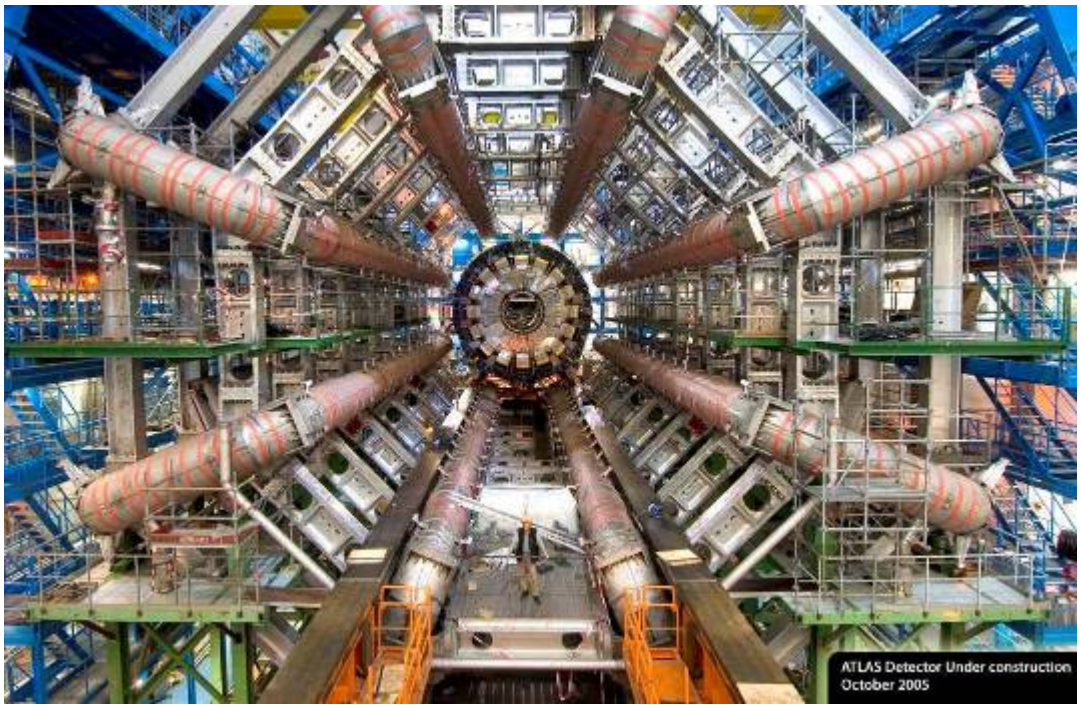


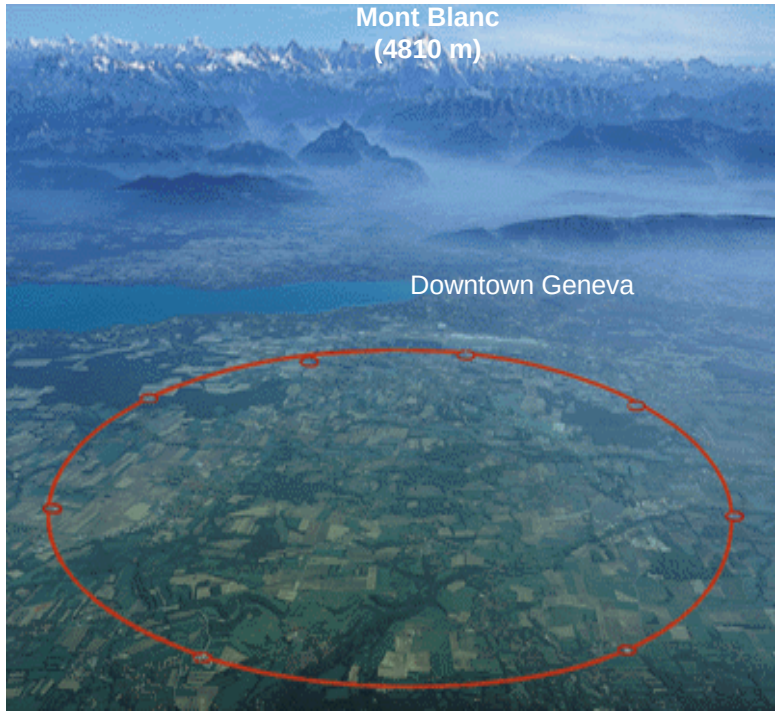
Book of abstracts from a recent User Forum:
http://doc.cern.ch//archive/electronic/egee/tr/egee-tr-2006-005.pdf

High Energy Physics

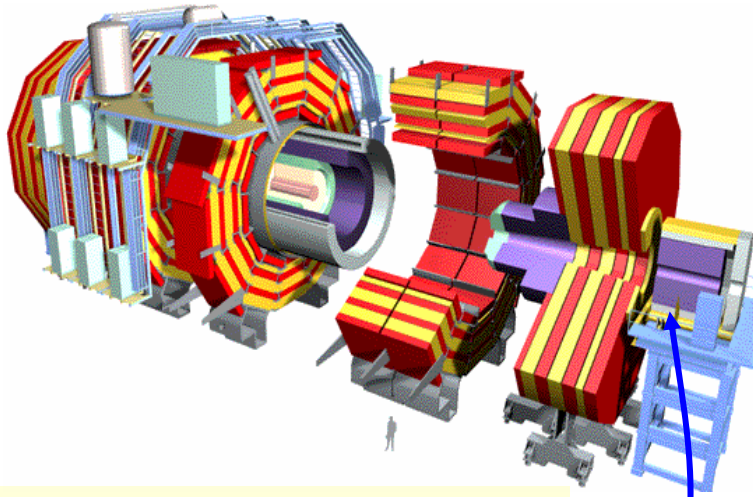
Large Hadron Collider (LHC):

- One of the most powerful instruments ever built to investigate matter
- 4 Experiments: ALICE, ATLAS, CMS, LHCb
- 27 km circumference tunnel
- Due to start up in 2007

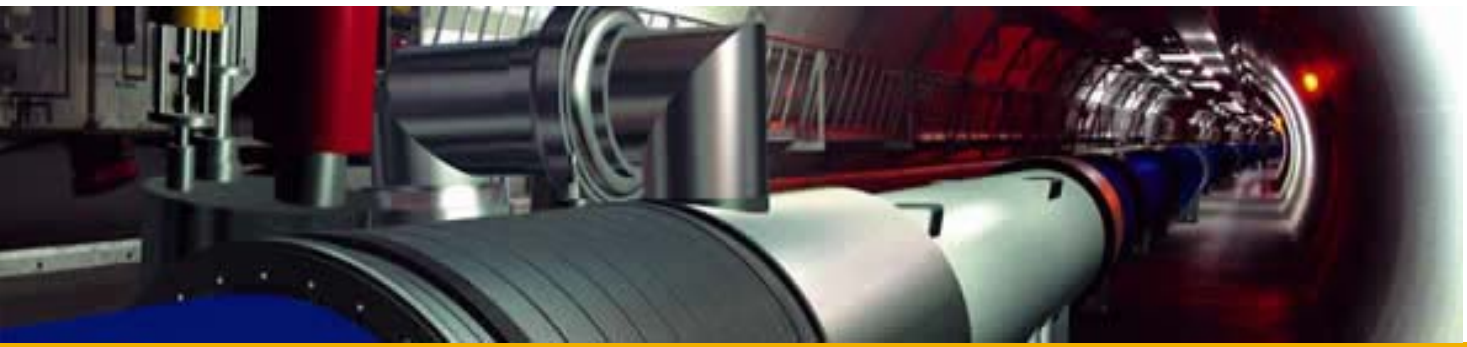




Large Hadron Collider data

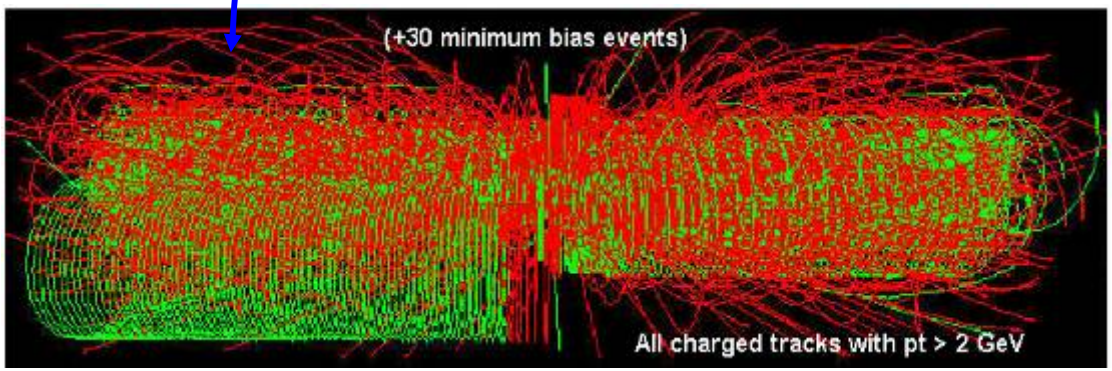


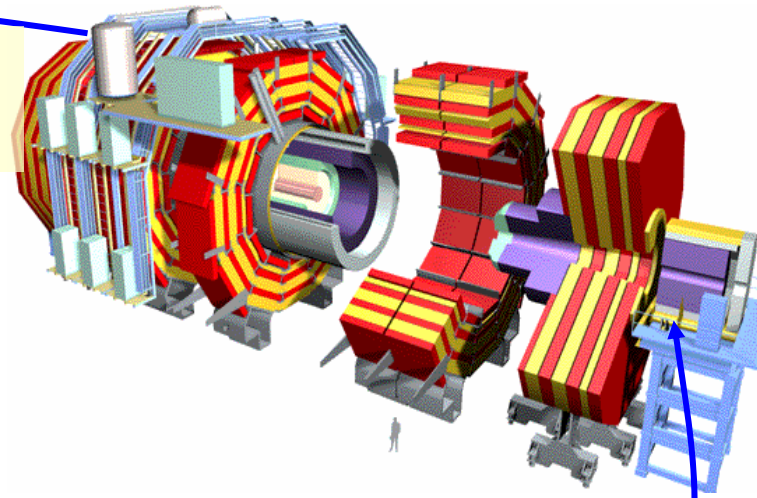
The accelerator generates 40 million particle collisions (events) every second at the centre of each of the four experiments' detectors



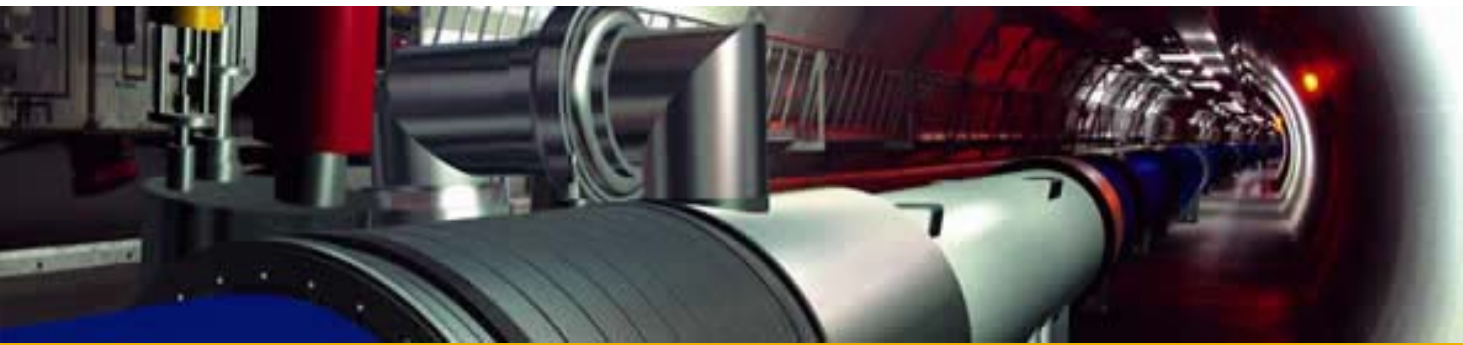
Large Hadron Collider data

This is reduced by online computers that filter out a few hundred “good” events/sec.





Which are recorded on disk and magnetic tape at 100-1,000 MegaBytes/sec → ~15 PetaBytes per year for all four experiments

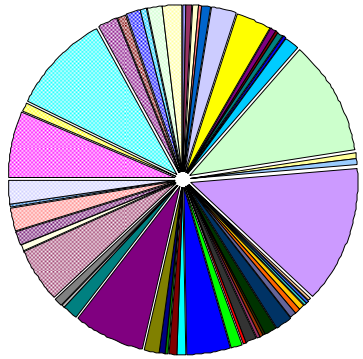


LHC data processing on the grid

- **LHC data and service challenges**
  - Preparing for LHC start-up in 2007
  - Ensure key services & infrastructure are in place
  - Emphasis on providing a service

LHCb

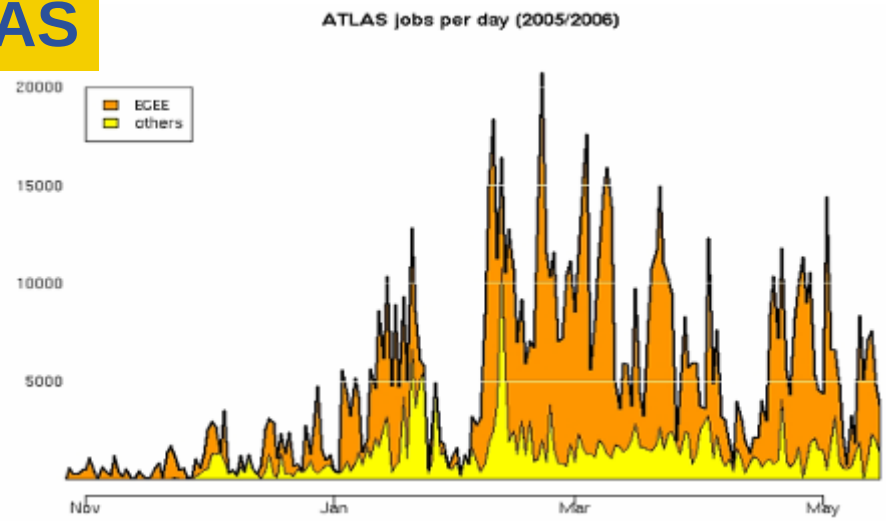
CPU used: 6,389,638 h
Data Output: 77 TB



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%

- **Computing needs of experiments**
  - E.g. LHCb: ~700 CPU years in 2005 on the EGEE infrastructure
  - E.g. ATLAS: over 10,000 jobs per day

ATLAS

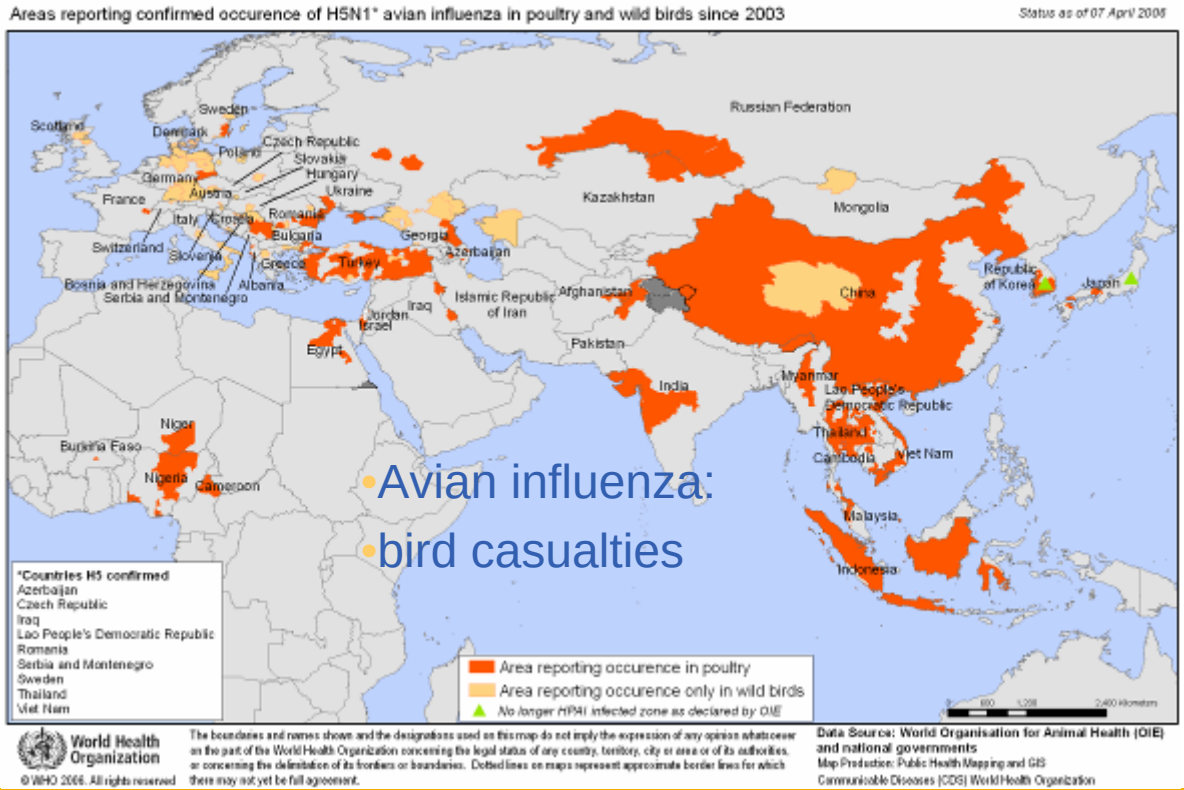


In silico drug discovery

- **Diseases such as HIV/AIDS, SRAS, Bird Flu etc. are a threat to public health due to world wide exchanges and circulation of persons**
- **Grids open new perspectives to *in silico* drug discovery**
  - Reduced cost and adding an accelerating factor in the search for new drugs

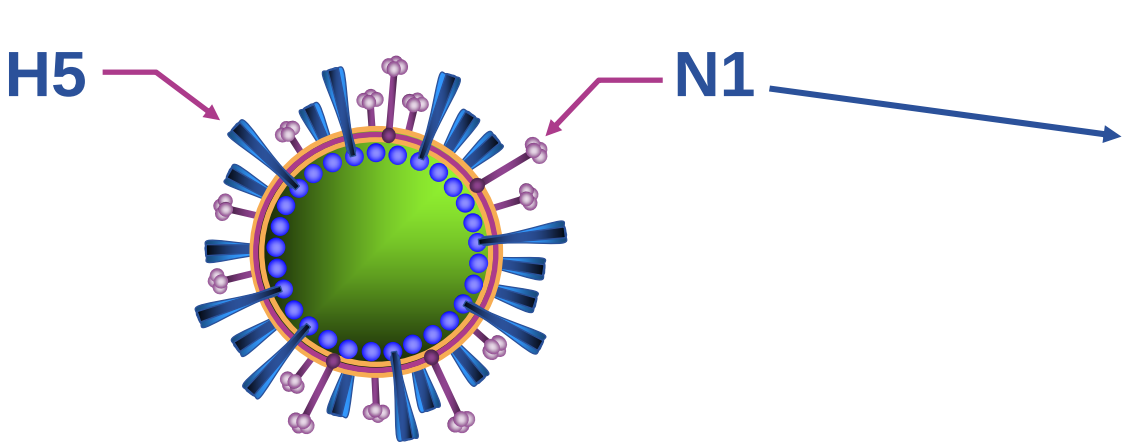
International collaboration is required for:

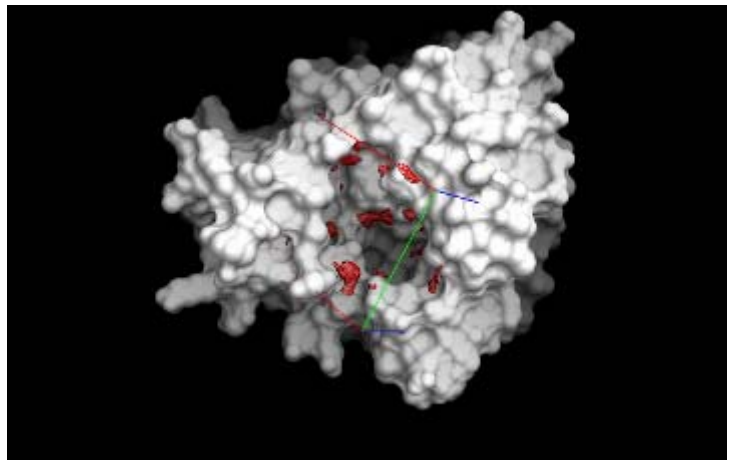
- **Early detection**
- **Epidemiological watch**
- **Prevention**
- **Search for new drugs**
- **Search for vaccines**



WISDOM, the first step

- **WISDOM focuses on drug discovery for neglected and emerging diseases.**
  - Summer 2005: World-wide Computation - In Silico Docking On Malaria
    - 46 million ligands docked in 6 weeks
      - *~1 million virtual ligands selected*
    - **1TB** of data produced
    - 1000 computers in 15 countries
      - *Equivalent to 80 CPU years*
  - Spring 2006: drug design against H5N1 neuraminidase involved in virus propagation
    - impact of selected point mutations on the efficiency of existing drugs
    - identification of new potential drugs acting on mutated N1



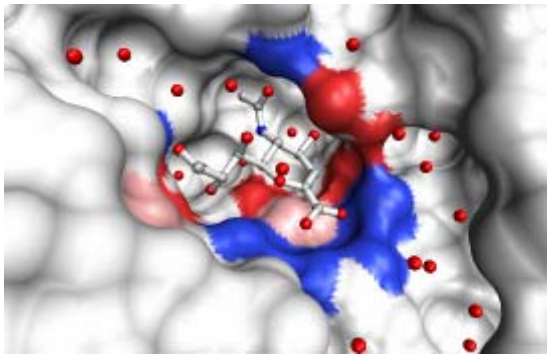


Challenges for high throughput virtual docking

Millions of chemical compounds available in laboratories

High Throughput Screening
2$/compound, nearly impossible

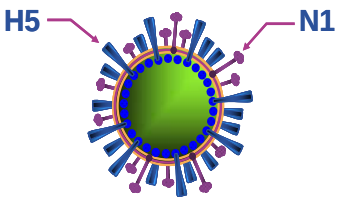
300,000 Chemical compounds:
ZINC &
Chemical combinatorial library



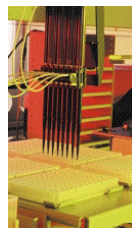
Molecular docking (Autodock)
~100 CPU years, 600 GB data

Data challenge on EGEE, Auvergrid, TWGrid
~6 weeks on ~2000 computers

Target (PDB) :
Neuraminidase (8 structures)



Hits sorting and refining



In vitro screening of 100 hits

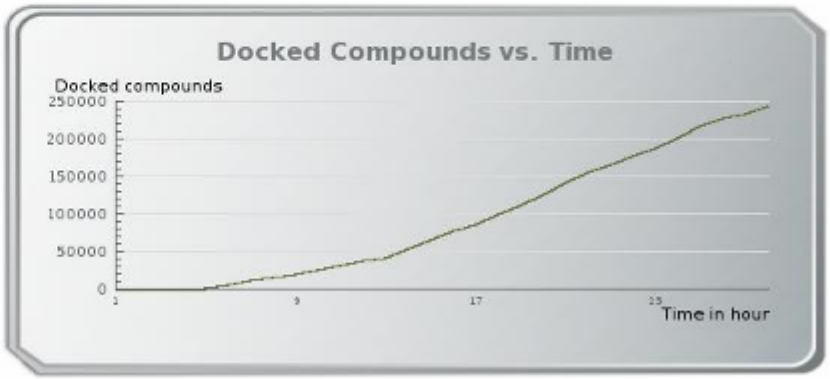
WISDOM

http://wisdom.healthgrid.org/

WISDOM

Initiative for grid-enabled drug discovery against neglected and emergent diseases

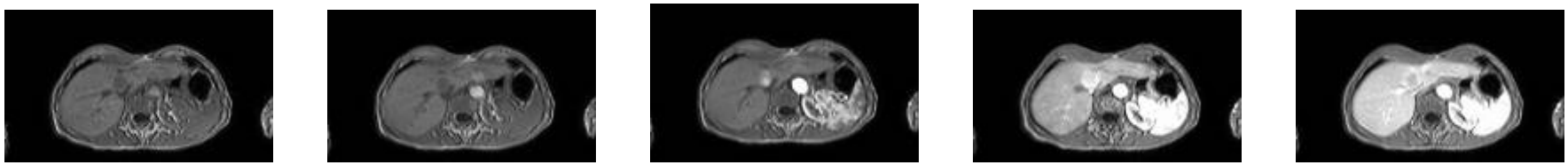
NUMBER OF DOCKED COMPOUNDS....	241200
IN SILICO COST.....	8.712 €
IN VITRO ESTIMATED COST.....	120.600 €
CPU.DAYS CONSUMED.....	363
SUCCESS RATE.....	83 %



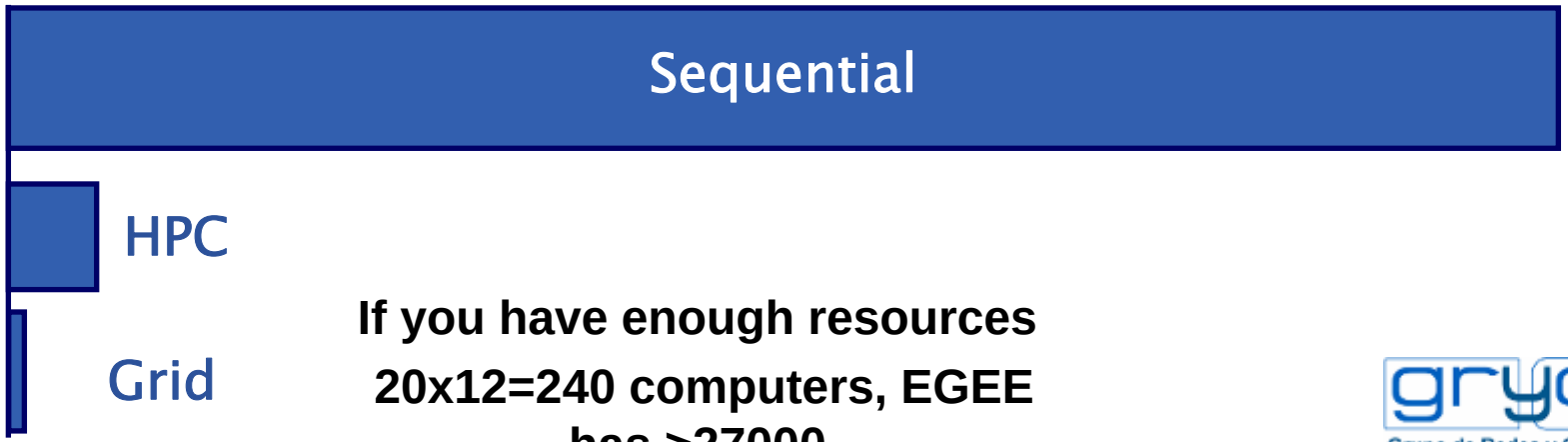
Docking challenge on EGEE and AuverGrid infrastructures

Medical image processing: analysing tumours

- **Pharmacokinetics: contrast agent diffusion study**
  - co-registration of a time series of volumetric medical images to analyse the evolution of the diffusion of contrast agents



- **Computational Costs**
  - 20 Patients: 2623 hours (Co-registration + Parametric Image)
  - Using a 20-processor Computing Farm: 146 hours
  - Using the Grid: <20 hours



If you have enough resources

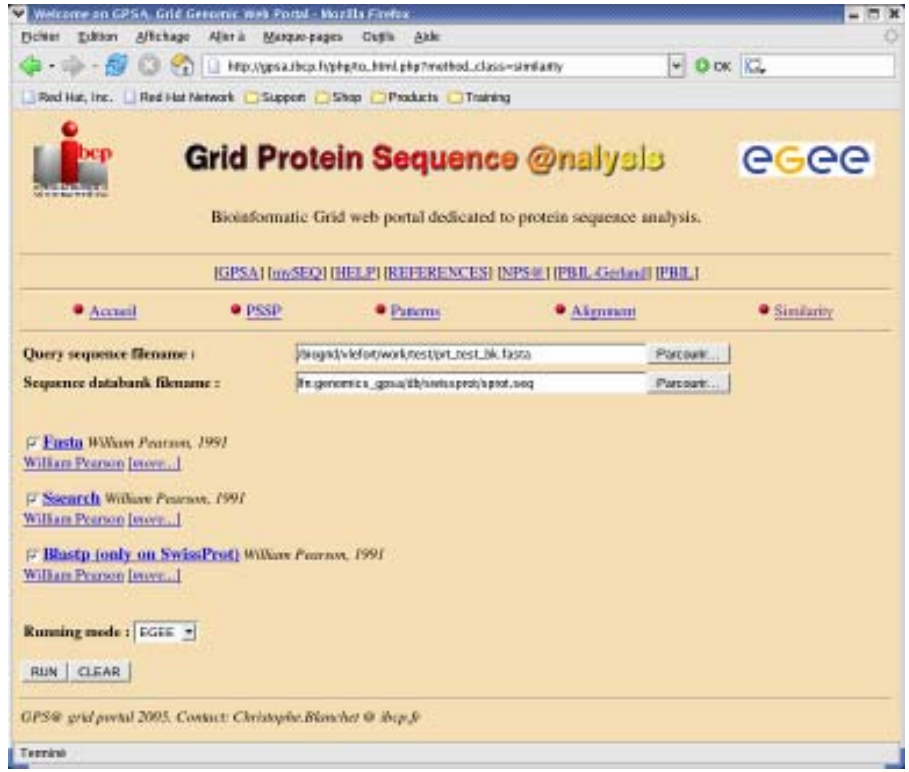
20x12=240 computers, EGEE has >27000

Grupo de Redes y Computación de Altas Prestaciones

Bioinformatics

GPS@: bioinformatics portal

- http://gpsa.ibcp.fr/ web portal
- Access up-to-date sequence and 3D-structure databanks (EMBL, GenBank, SWISS-PROT etc.)
- Tens of bioinformatics legacy code



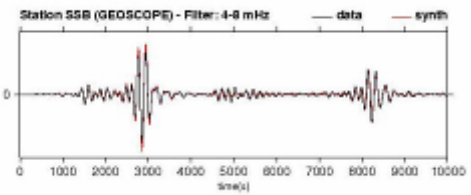
- **Convenient easy-to-use interface with access to well-known databanks**
- **Uses grid resources to analyse the sequences**

Example: Determining earthquake mechanisms

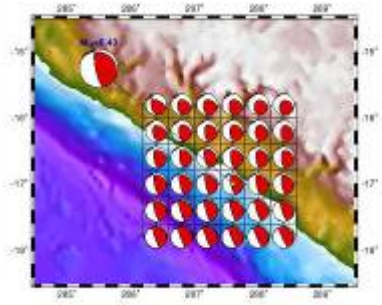
- **Seismic software application determines epicentre, magnitude, mechanism**

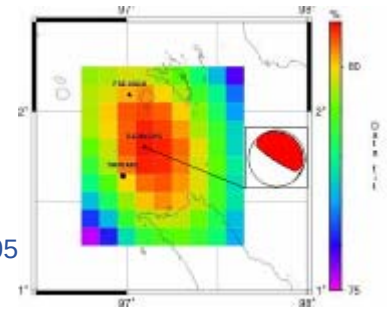
- **Analysis of Indonesian earthquake (28 March 2005)**
  - Seismic data within 12 hours after the earthquake
  - Analysis performed within 30 hours after earthquake occurred
    - 10 times faster on the Grid than on local computers
  - Results
    - Not an aftershock of December 2004 earthquake
    - Different location (different part of fault line further south)
    - Different mechanism

→ **Rapid analysis of earthquakes important for relief efforts**



Peru, June 23, 2001
Mw=8.4

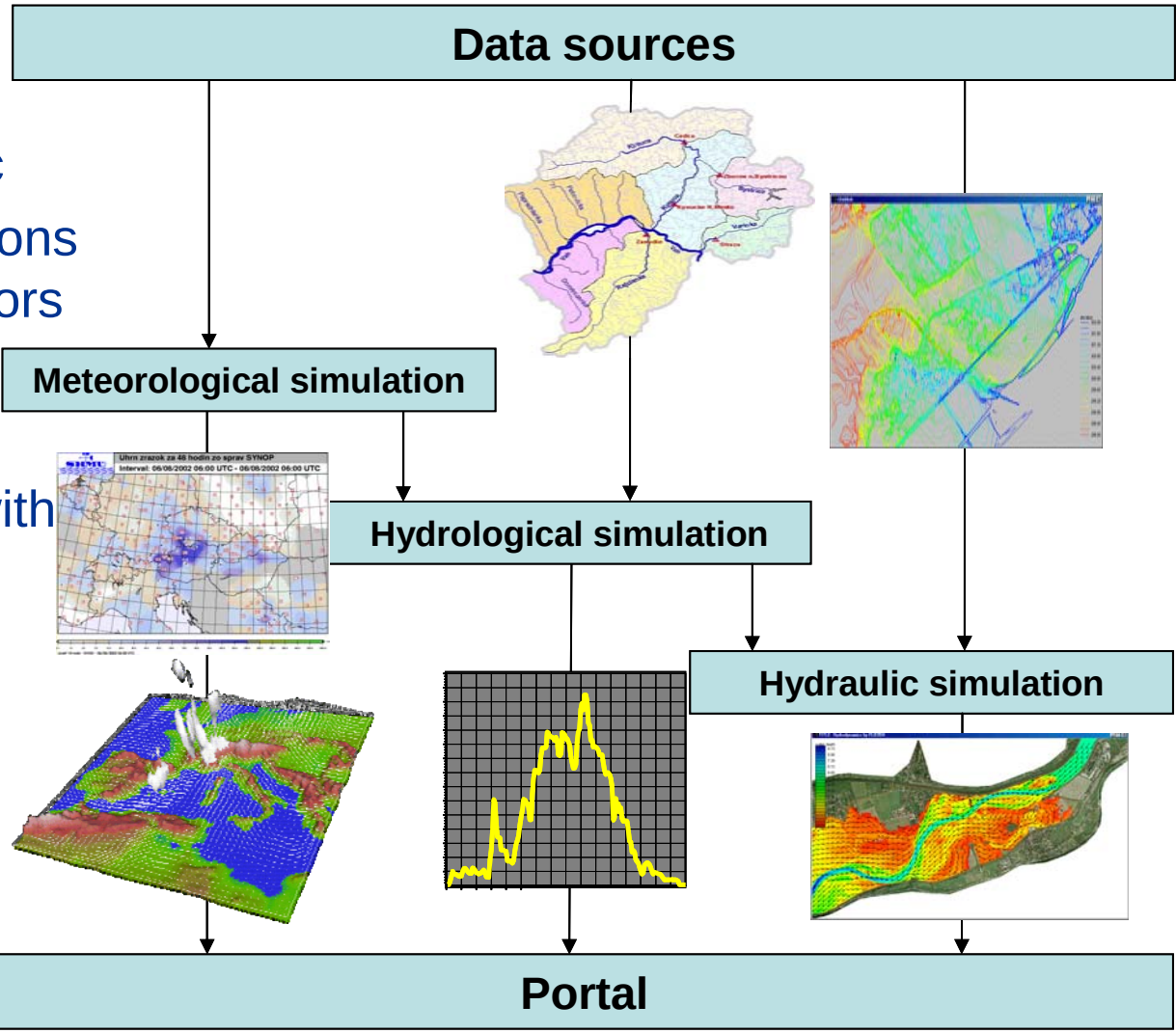




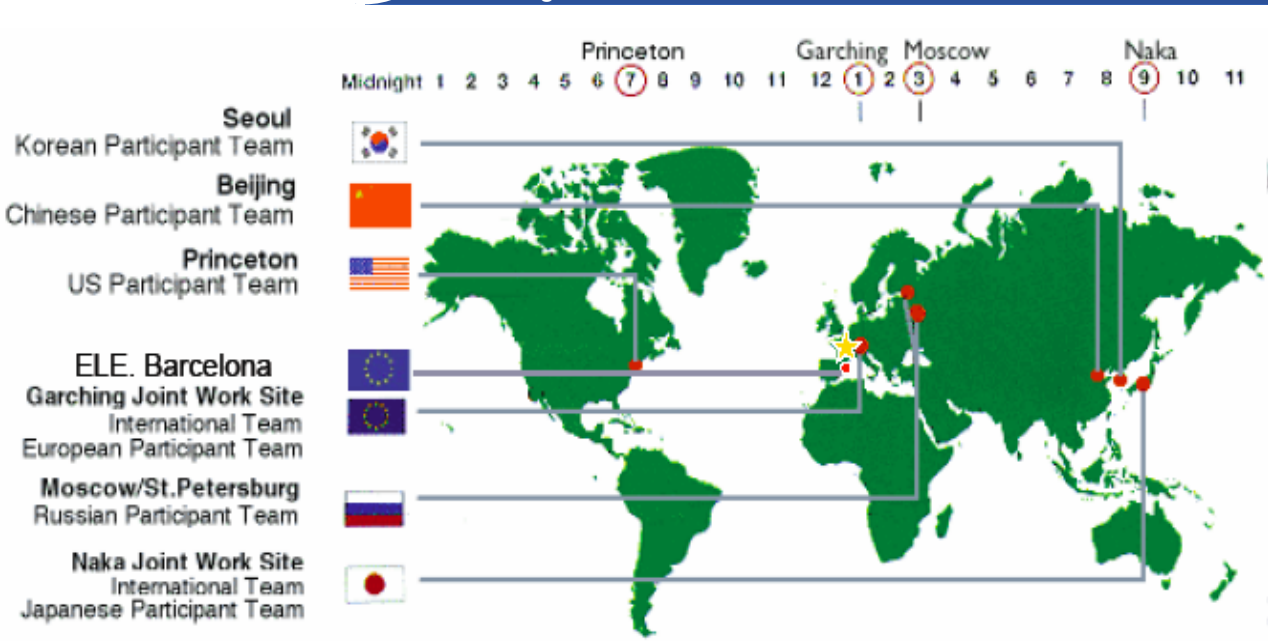
Sumatra, March 28, 2005
Mw=8.5

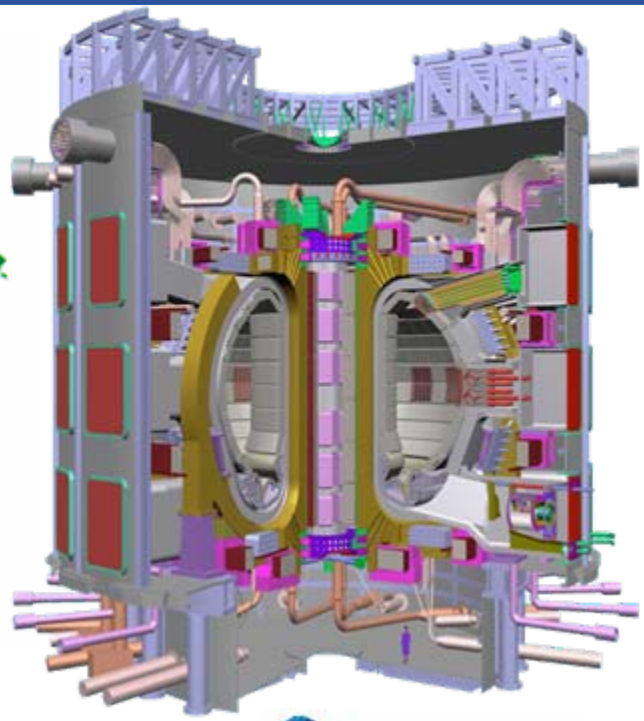
Flood forecasting problem

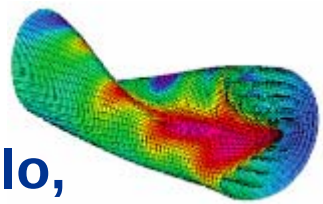
- **Many kinds of data**
  - Meteorological, hydrological, hydraulic
  - Generated by simulations or obtained from sensors
  - Permanent or periodically updated
  - Publicly available or with restricted access



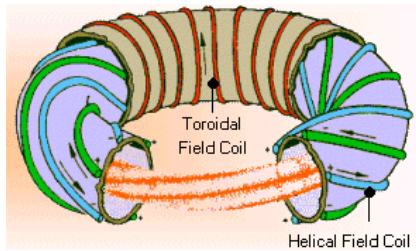
Future ITER Fusion reactor





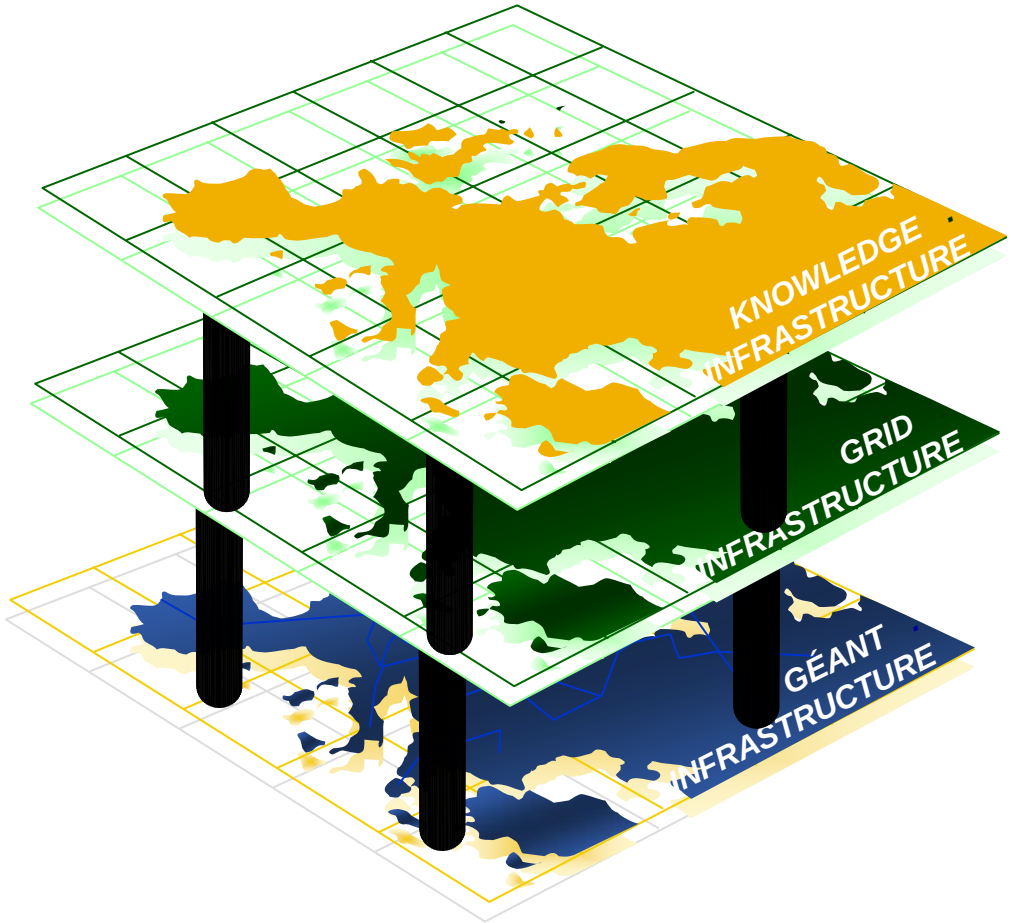


- Applications with distributed calculations: Monte Carlo, Separate estimates, …
- Multiple Ray Tracing: e. g. TRUBA
- Stellarator Optimization: VMEC
- Transport and Kinetic Theory: Monte Carlo Codes



Support for digital repositories

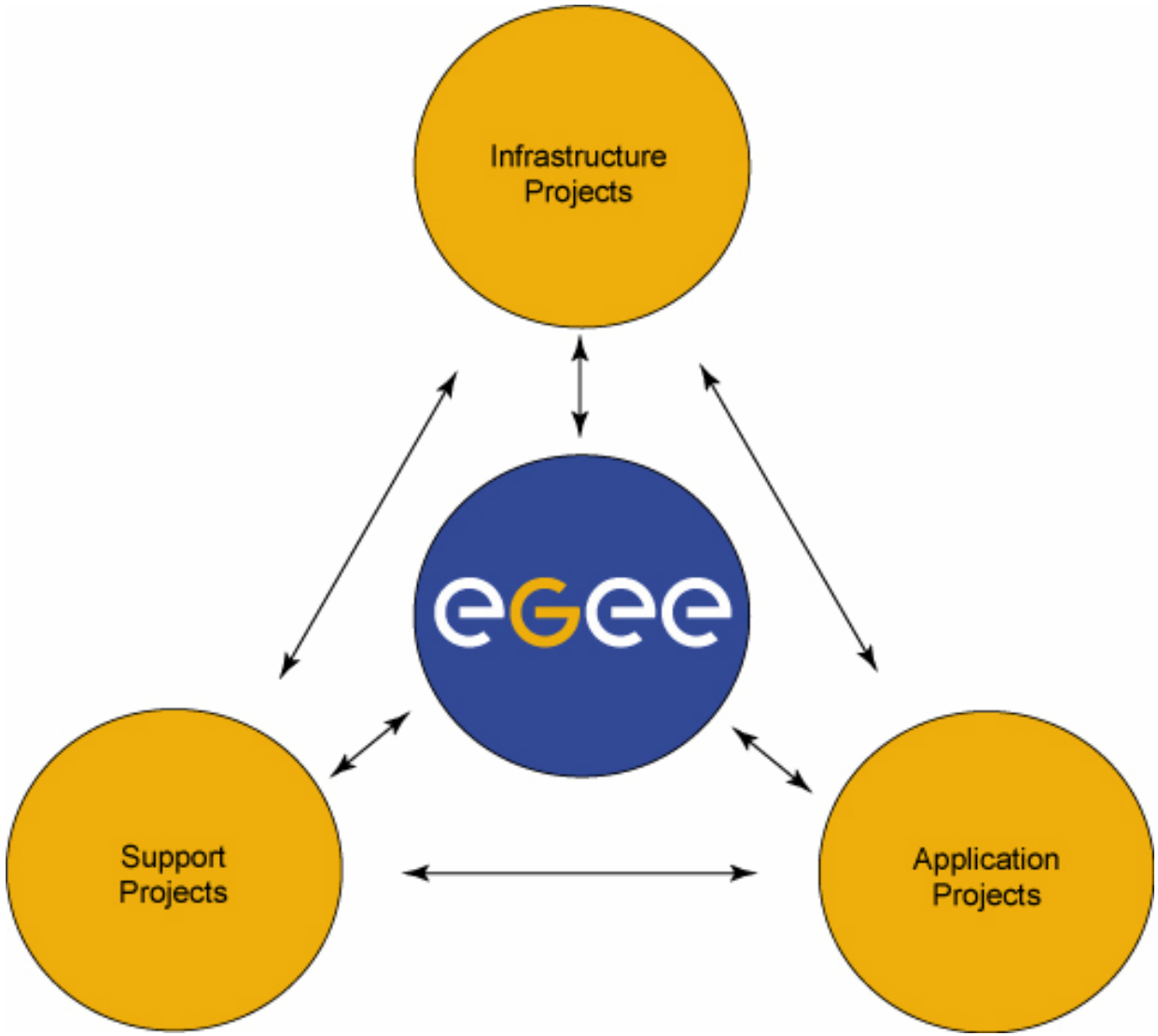
3 layered model to support access to heterogeneous information and connect resources through common shared services



Grids can offer:

- Sharing of resources
- Secure Access Control
- Data management
- Execution of computationally demanding applications (e.g. multi-media content)

Projects related to EGEE



Summary

- **Grids are all about sharing in Virtual Organisations – communities spread throughout the world sharing computing resources, data, software, information in a cooperative environment**
- **Inter-operability is key to providing the level of support required for our user communities**
- **EGEE Infrastructure – world's largest multi-science production grid service**
- **EGEE-II is the opportunity to expand on this existing base both in terms of scale and usage**
- **EGEE has already put in place a support structure for many applications and is working with more scientific communities to further extend grid usage**
- **Need to prepare the long-term**
  - EGEE, related EU projects, national grid initiatives(NGIs) and user communities are working to define a model for a sustainable grid infrastructure that is independent of short project cycles

www.eu-egee.org

Learning more and getting involved in grids

- **Trying things out**
  - http://www.eu-egee.org/try-the-grid
  - 'under construction'!
  - Comments /complaints to f.harris@cern.ch, hannelore.hammerle@cern.ch, egee2@metaware.it
- **Another talk on grid in summer school series (more from HEP and middleware point of view) – P. Mendes Lorenzo, Aug 7**
  - http://agenda.cern.ch/fullAgenda.php?ida=a062808
- **Book of abstracts from a recent User Forum (a very broad range of applications and grid experiences/issues):**
  - http://doc.cern.ch//archive/electronic/egee/tr/egee-tr-2006-005.pdf
- **EGEE training events (planned all over Europe on broad range of topics)**
  - http://www.egee.nesc.ac.uk/index.html
- **'The grid – blueprint for a new computing infrastructure', Foster and Kesselman, 1998, ISBN 1-55860-475-8**
- **Links to possible job opportunities**

  http://egee-technical.web.cern.ch/egee-technical/jobs/jobs.htm

Some interesting links

- **The EGEE Project**
  - http://www.eu-egee.org
- **The ICEAGE Project (an EU education project)**
  - http://www.**iceage**-eu.org
- **The LCG Project (LHC Computing Grid)**
  - http://cern.ch/lcg
- **The gLite middleware (EGEE middleware)**
  - http://www.glite.org
- **The Condor Project (more on middleware for grid resource management)**
  - http://www.cs.wisc.edu/condor
- **The Globus Project (the original source of basic grid middleware)**
  - http://www.globus.org
- **Website for international grid forum**
  - http://www.gridforum.org /

- **QUESTIONS……..**