



Enabling Grids for E-science in Europe

The EGEE Grid infrastructure project: first experience and future plans in the area of life sciences

By Vincent Breton

EGEE application coordinator

CNRS

Clermont-Ferrand

France

- Goal: empower biological analysis workflow on up-to-date exponentially growing data
- Technical challenges
 - data and tools integration : address data heterogeneity and legacy of tools and standards
 - provide the infrastructure and the services (database update and mirroring, grid portals, toolboxes)
- Human challenge : involve end users in the grid game
 - Grids are still very much in development and therefore user-unfriendly
 - Training and support to molecular biology tool and data providers, university hospitals, biology/medicine research centres, ...

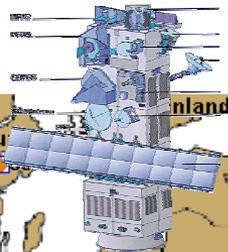
- Goal: allow every physician to access a reliable grid for his daily practice
 - New actors : hospitals, physicians, healthcare administrations
- Technical challenges
 - Networking, User interface
 - Grid quality of services (stability, scalability, security, privacy, ...)
 - Legal/ethical issues: obey the laws of International countries with respect to personal data ownership and data transfer
- Human challenge: new approach to healthcare delivery
 - Change the way in which doctors/healthcare administrations conceive health
 - Grids are still very much in development and therefore user-unfriendly
 - Training and support to healthcare professionals

- Several European projects within the 5th framework program
 - Multidisciplinary projects: CrossGrid, DataGrid, EuroGrid, ...
 - Projects focused on healthcare issues: GEMSS, Mammogrid, ...
- International/National projects
 - UK e-science (Mygrid, ...)
 - French ACI grid (Medigrid, GRIPPS, GLOP, ...)
 - American projects (BIRN, NDMA, ...)
 - Japanese projects (Biogrid, ...)
 - Asia-Pacific (APBiogrid, ...)
 - Russian projects (RGrid, ...)
 - ...

Centres for E-science in Europe

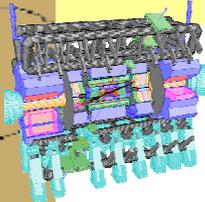
People

- 500 registered users
- 12 Virtual Organisations
- 21 Certificate Authorities
- >600 people trained
- 456 person-years of effort
- 170 years funded



Application Testbed

- ~20 regular sites
- > 60,000 jobs submitted (since 09/03, release 2.0)
- Peak > 1000 CPUs
- 6 Mass Storage Systems



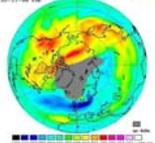
Coordinator:
F. Gagliardi,
CERN

Software

- > 65 use cases
- 7 major software releases (> 60 in total)
- > 1,000,000 lines of code

Scientific Applications

- 5 Earth Obs institutes
- 10 bio-medical apps
- 6 HEP experiments



- DataGrid involved three research communities
 - High Energy Physics
 - Life sciences and medical imaging
 - Earth observation

- Obvious differences in communities organization and computing awareness had to be addressed
 - from the biology community which has no center of gravity and where most end-users are not skilled at using computers...
 - to High Energy Physics community which is extremely organized around CERN with skilled users
 - Setting-up multidisciplinary coordination at application level turned out very beneficial

- Most user communities are not willing to play guinea pigs of grid technology
 - Added value to be demonstrated early on to biologists

- Cultural: difficulty to establish a common language between the project partners
- Cultural: Need for intermediate levels between middleware developers and “end”-users
 - Application developers aware of middleware issues needed to act as an interface with end-users (biologists, physicists, physicians)
- Social: a grid deployment project brings a real sense of identity
 - DataGrid ended up as a true collaboration with a team spirit, to the benefit of EGEE

- | | |
|---|---------------|
| ■ | deployed |
| ■ | tested on EDG |
| ■ | Not achieved |

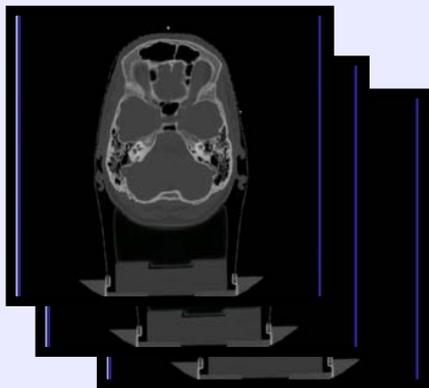
- **Bio-informatics**

- **Phylogenetics** : BBE Lyon (T. Sylvestre)
- **Search for primers** : Centrale Paris (K. Kurata)
- **Bio-informatics web portal** : IBCP (C. Blanchet)
- **Parasitology** : LBP Clermont, Univ B. Pascal (N. Jacq)
- **Data-mining on DNA chips** : Karolinska (R. Médina, R. Martinez)
- **Geometrical protein comparison** : Univ. Padova (C. Ferrari)

- **Medical imaging**

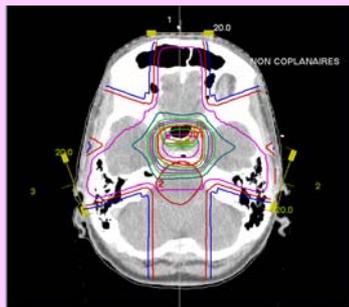
- **MR image simulation** : CREATIS (H. Benoit-Cattin)
- **Medical data and metadata management** : CREATIS (J. Montagnat)
- **Mammographies analysis** ERIC/Lyon 2 (S. Miguet, T. Tweed)
- **Simulation platform for PET/SPECT based on Geant4** : GATE collaboration (L. Maigne)

1°) Obtain scanner slices images



The head is imaged using a MRI and/or CT scanner

2°) Treatment planning



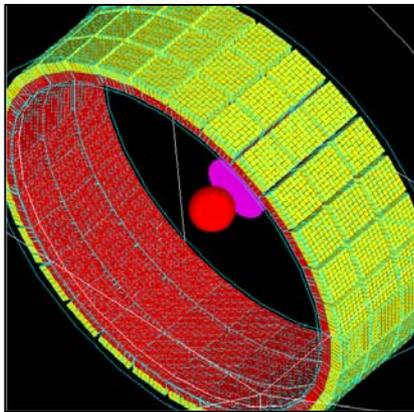
Calculation of deposit dose on the tumor (~1mn):
A treatment plan is developed using the images

3°) Radiotherapy treatment

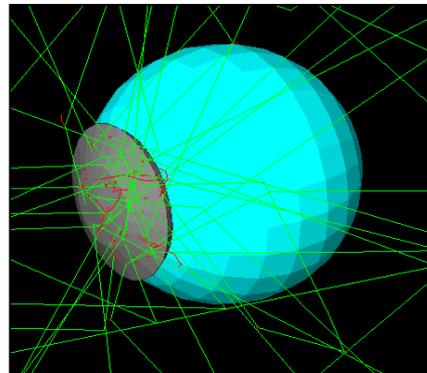


Irradiation of the brain tumor with a linear accelerator

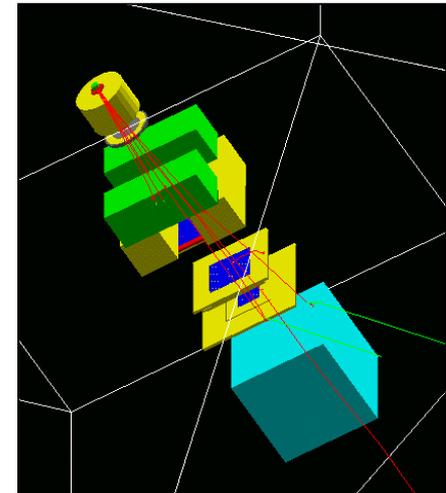
- Today: analytic calculation to compute dose distributions in the tumor
 - For new Intensity Modulated Radiotherapy treatments, analytic calculations off by 10 to 20% near heterogeneities
- Alternative: Monte Carlo (MC) simulations in medical applications
- The GRID impact: reduce MC computing time to a few minutes



PET camera

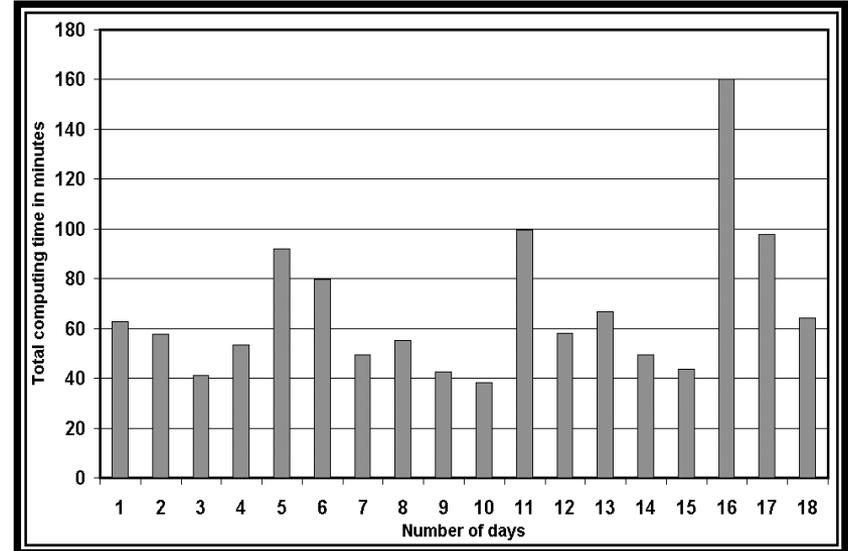
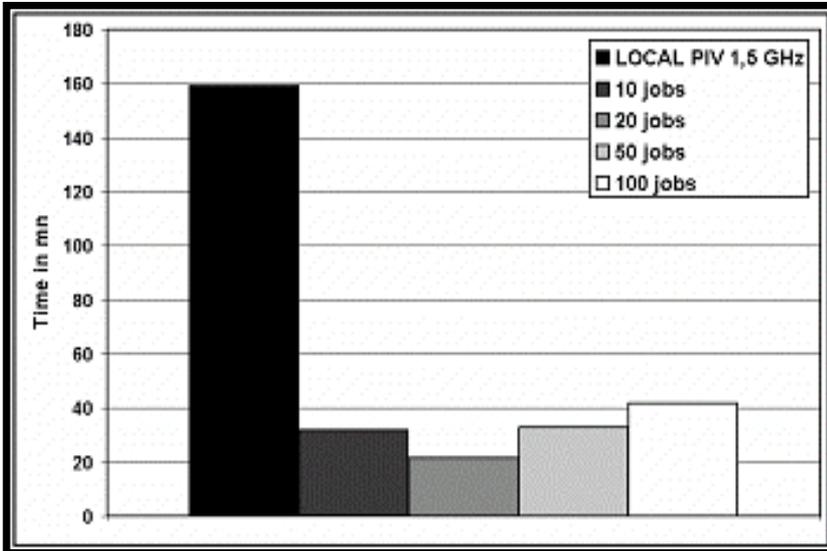


Ocular brachytherapy treatment



Radiotherapy

- The parallelization of GATE on the DataGrid testbed has shown significant reduction in computing time (factor 10)

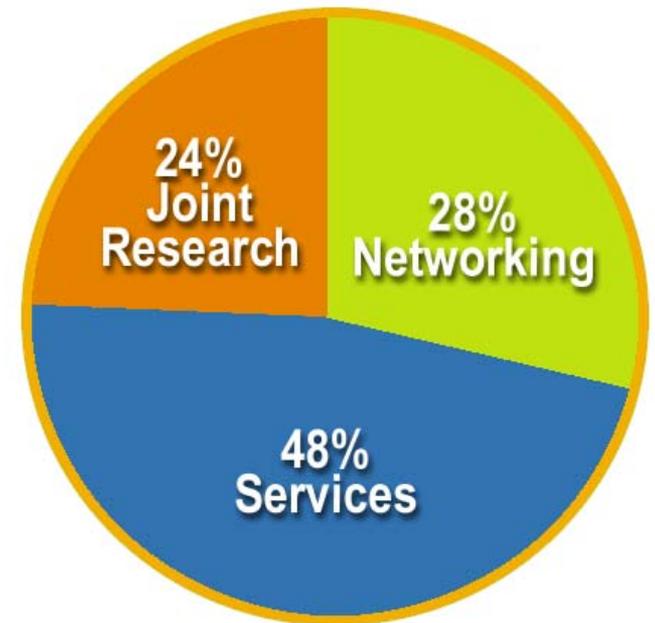


- Issue for clinical routine: how to guarantee grid response time ?
 - Job prioritization on PC farms
 - Job submission to a supercomputer
 - Reduce grid services response time

- 70 leading institutions in 27 countries, federated in regional Grids
- 32 M Euros EU funding (2004-5), O(100 M) total budget
- Aiming for a combined capacity of over 20'000 CPUs (one of the largest international Grid infrastructures ever assembled)
- ~ 300 dedicated staff



- Emphasis on operating a production grid and supporting the end-users
- **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)



- Middleware selected based on requirements of Applications and Operations
- Harden and re-engineer existing middleware functionality, leveraging the experience of partners
- Provide robust, supportable components
- Support components evolution towards a service oriented approach (Web Services)



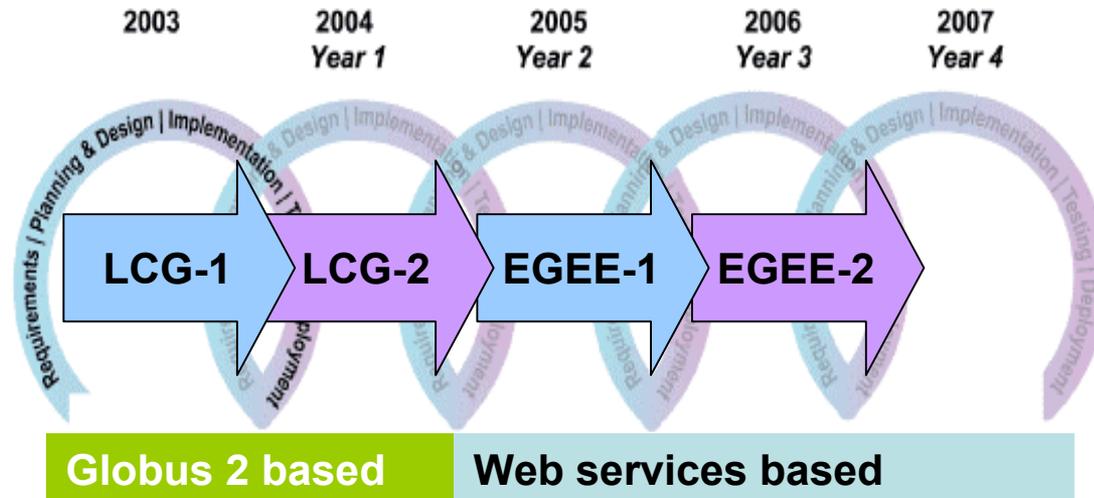
- Middleware Integration and Testing Centre
- Middleware Re-engineering Centre
- Quality and Security Centres

- **gLite**

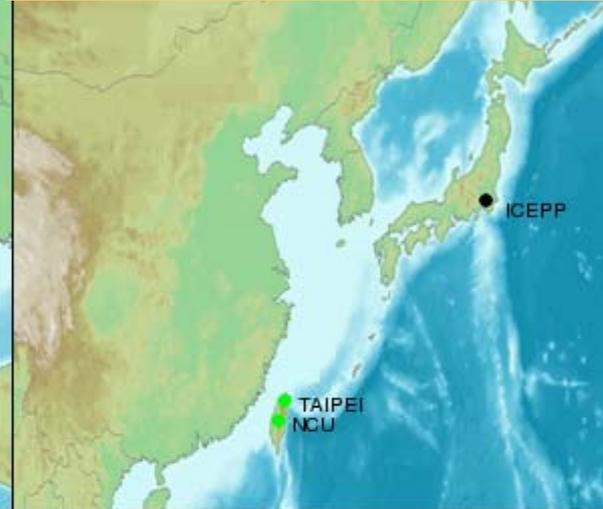
- Exploit **experience and existing components** from VDT (CondorG, Globus), EDG/LCG, AliEn, and others
- Develop a **lightweight stack of generic middleware** useful to EGEE applications (HEP and Biomedics are pilot applications).
 - Should eventually deploy dynamically (e.g. as a globus job)
 - Pluggable components – cater for different implementations
- Focus is on **re-engineering and hardening**
- Early **prototype** and fast feedback turnaround envisaged



- **From day 1 (1st April 2004)**
 - Production grid service based on the LCG infrastructure running LCG-2 grid middleware (SA)
 - LCG-2 will be maintained until the new generation has proven itself (fallback solution)
- **In parallel develop a “next generation” grid facility**
 - Produce a new set of grid services according to evolving standards (Web Services)
 - Run a development service providing early access for evaluation purposes
 - Will replace LCG-2 on production facility in 2005



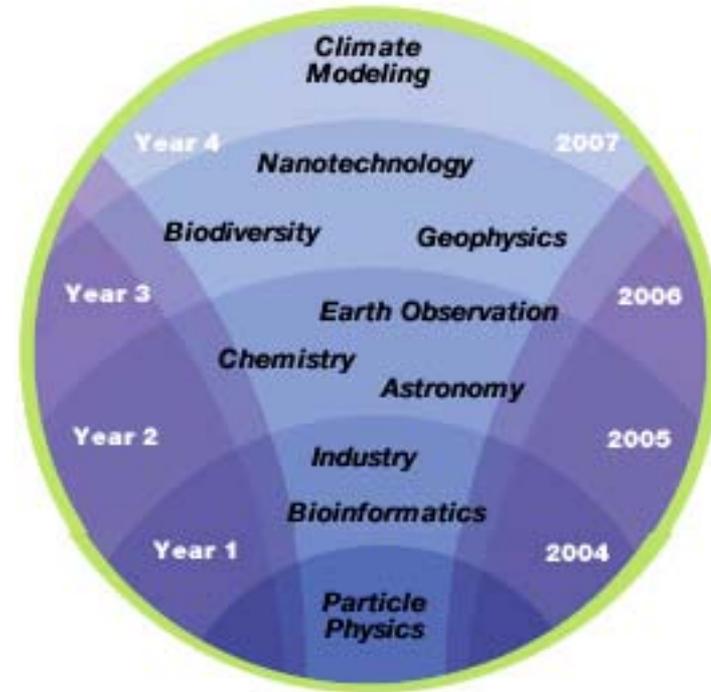
- 22 Countries
- 58 Sites (45 Europe, 2 US, 5 Canada, 5 Asia, 1 HP)
 - Coming: New Zealand, China, other HP (Brazil, Singapore)
- 3800 cpu



- No information
- Scheduled Maintenance
- Timeout
- Globus OK
- RB OK

Status for Resource Broker CERN_lxn1188: Thu Jun 3 16:45:34 BST 2004

- EGEE Scope : ALL-Inclusive for academic applications (open to industrial and socio-economic world as well)
- The major success criterion of EGEE: how many satisfied users from how many different domains ?
- Goal: 5000 users (3000 after year 2) from at least 5 disciplines



Application domains and timelines are for illustration only

- 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 application areas – Particle Physics and Life sciences, while developing a process for supporting other application areas

- Two scientific areas selected to guide the implementation and certify the performance and functionality of the evolving infrastructure: **Particle Physics & Life sciences**
- Physics pilot applications are LHC experiments (ALICE, ATLAS, CMS, LHCb)
- Life sciences pilot applications
 - GPS@: web portal for bioinformatics
 - GATE: Monte-Carlo platform for radiotherapy treatment planning
 - CDSS: expert system for medicine

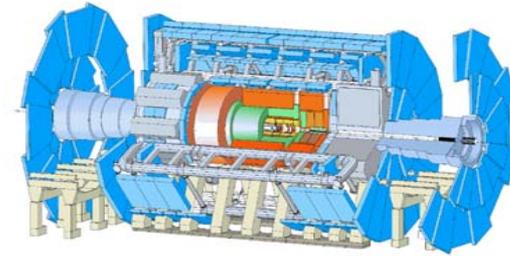
HEP:

- Running large distributed computing systems for many years
- Focus for the future is on computing for LHC (LCG)
- The 4 LHC experiments and other current HEP experiments use grid technology e.g. Babar,CDF,D0..,
- LHC experiments are currently executing large scale data challenges(DCs) involving thousands of processors world-wide and generating many Terabytes of data
- Moving to so-called ‘chaotic’ use of grid with individual user analysis (thousands of users interactively operating within experiment VOs)

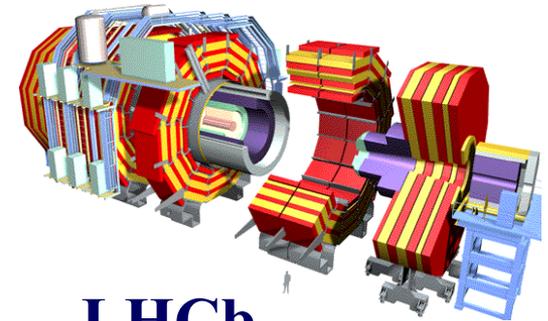


- **Storage**
 - Raw recording rate 0.1 – 1 GByte/s
 - Accumulating at 5-8 PetaByte/year
 - 10 PetaByte of disk
- **Processing**
 - 200,000 of today's fastest PCs

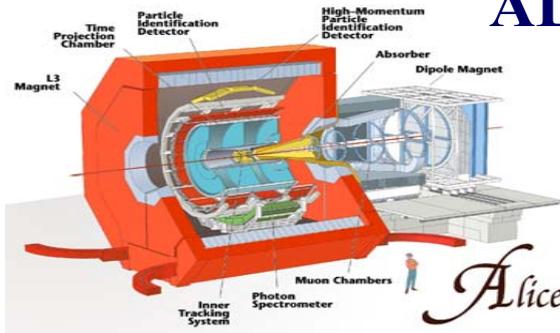
ATLAS



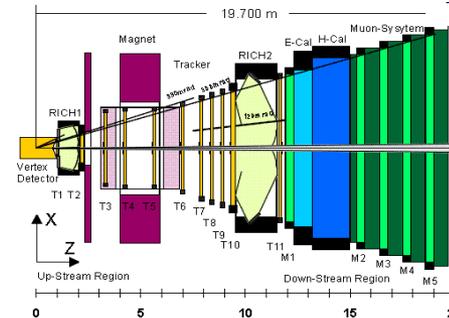
CMS



ALICE



LHCb



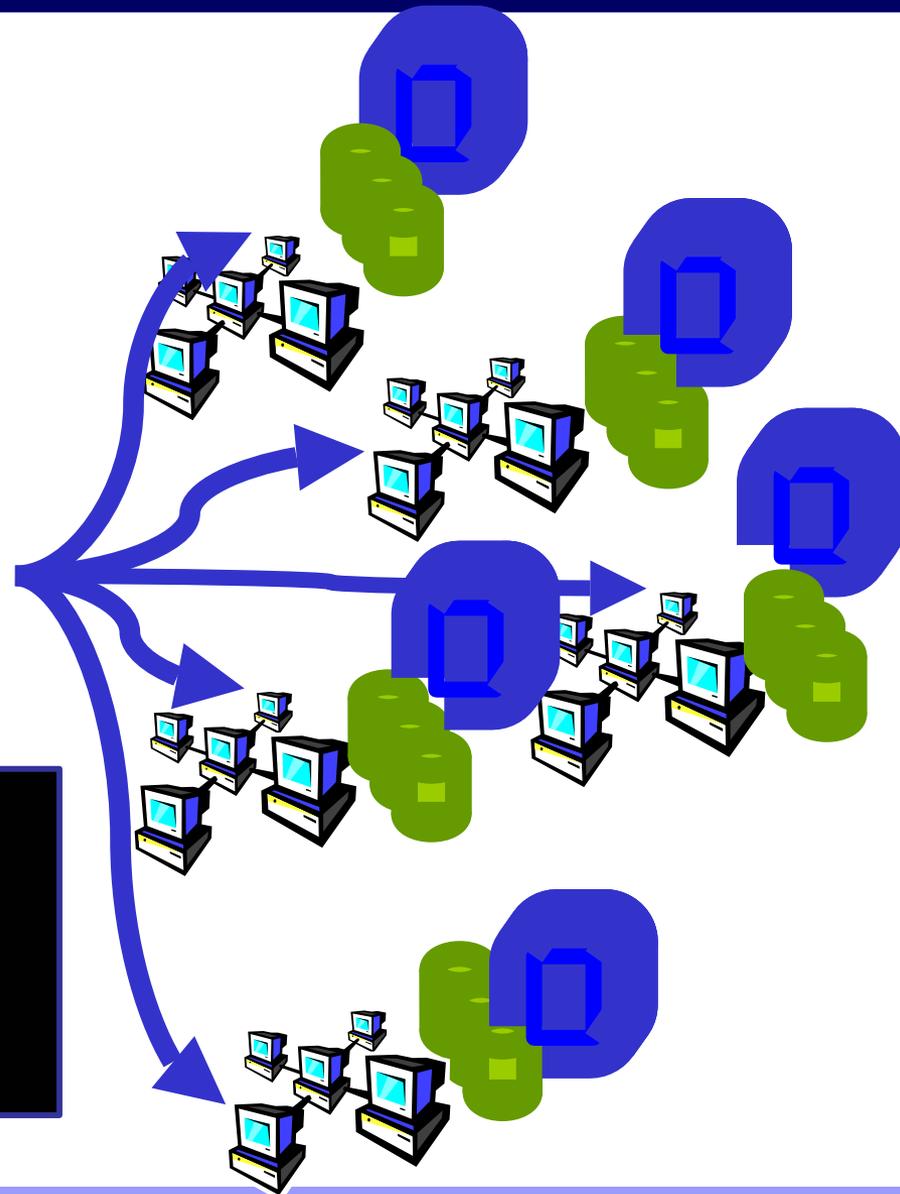
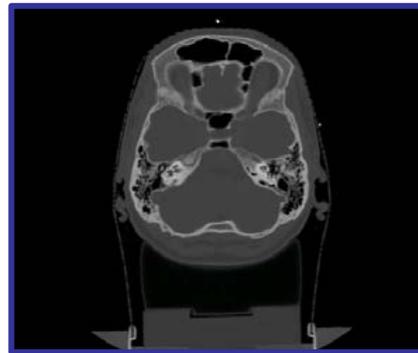
- Goals

- To deploy biomedical applications on EGEE infrastructure
- To feed back requirements to the middleware developers
 - Requirements database: <http://egee-na4.ct.infn.it/requirements/>
- To participate to the early testing of EGEE middleware
- To set up virtual organizations and to integrate new resources into the EGEE infrastructure
- To establish collaborations with (national, European, Worldwide) life science projects for deployment on EGEE

- Partners

- CNRS, Universidad Politecnica de Valencia, CSIC (CNB Madrid)

- **Contact: Lydia Maigne,**
maigne@clermont.in2p3.fr
- **Goal: address**
radiotherapy treatment
planning on a grid
- **Deployment and status:**
installed on 2 EGEE
nodes (Clermont-Ferrand,
Lyon)
- **Main challenges:**
guaranteed response time



- Contact: Christophe Blanchet, christophe.blanchet@ibcp.fr
- Description: Web portal for bioinformatics
- Deployment and status
 - NPSA is a production web portal hosting proteins databases and algorithms
 - GPS@ is the grid version under development deployed on LCG2
- Users
 - NPSA serves hundreds of bioinformaticians daily (about 3000 jobs/day) but limited resources (4 CPUs)
- Plans
 - to replace NPSA with GPS@ when showing similar robustness
- Main challenge: reduce response time for short jobs

- Contact: Ignacio Blanquer, iblanque@dsic.upv.es
- Description: Clinical Decision Support System: expert system for medicine
- Deployment and status: original developed under serviced-based approach, now ported to EGEE-0
- Users
 - About 10 medical users from 5 organizations
 - About 10 runs per day (1 hour each)



- Beside "pilot" applications used to test EGEE middleware and to evaluate performances
- "internal" applications
 - come from within the project in the sense that they involve EGEE partners in collaboration with institutes external to EGEE (ex Babar, UK e-science projects,...)
 - have already a good middleware experience.
 - should be identified as they are often deployed at a national level and are therefore extremely dependent on interoperability between EGEE and national initiatives.
- "external" applications
 - from collaborations external to EGEE (european projects, national projects,...)
 - .

A new scientific community makes first contacts to EGEE through outreach events organized by Networking Activities

Follow-up meetings by applications specialists may lead to definition of new requirements for the infrastructure



Peer communication and dissemination events featuring established users then attract new communities



Implementation



If approved, the requirements are implemented by the Middleware Activities



The Networking Activities then provide appropriate training to the community in question, so that it becomes an established user

After integration and testing, the new middleware is deployed by the Service Activities

- Training material and courses from introductory to advanced level developed at NeSC in UK
- Train a wide variety of users both internal to the EGEE consortium and external groups from around the world
- 12 courses/presentations already held many more planned in the future
- Experience with GENIUS portal and GILDA testbed (provided by INFN)
- Major participation to second International Grid school in Italy



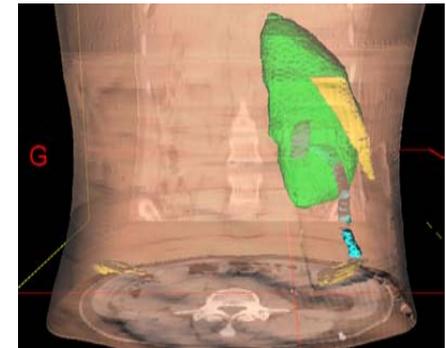
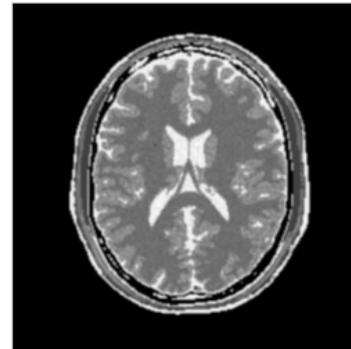
- Getting new scientific and industrial communities interested and committed to use the grid infrastructure built by EGEE is key to the success of the project
- Questionnaire to get information and first requirements from new communities interested in using the EGEE Infrastructure (<http://alipc1.ct.infn.it/grid/egee/na4/questionnaire/na4-genapp-questionnaire.doc>)
- Feed-backs received so far (<http://alipc1.ct.infn.it/grid/egee/na4/questionnaire>):
 - Astrophysics (EVO and Planck satellite)
 - Earth Observation (ozone maps, seismology, climate)
 - Digital Libraries (DILIGENT Project)
 - Grid Search Engines (GRACE Project)
 - Industrial applications (SIMDAT Project)
- Interest also from Computational Chemistry (Italy and Czech Republic), Civil Engineering (Spain), and Geophysics (Switzerland and France) communities

- 0 Review information provided on the EGEE website (www.eu-egee.org)
- 1 Establish contact with the EGEE applications group
 - Life sciences: Johan Montagnat (Johan.Montagnat@CREATIS.INSA-LYON.FR)
 - “Generic” (non HEP and Life Sciences) applications: Roberto Barbera (Roberto.Barbera@ct.infn.it)
- 2 Provide information by completing a questionnaire describing your application
- 3 Applications are selected for direct support based on scientific criteria, Grid added value, effort involved in deployment, resources consumed/contributed etc.
- 4 Follow a training session
- 5 Migrate application to EGEE infrastructure with the support of EGEE experts
- 6 Initial deployment for testing purposes
- 7 Production usage
 - Contribute computing resources for heavy production demands

- Name: SiMRI3D
- Contact: Fabrice Bellet, fabrice.bellet@creatis.insa-lyon.fr
- Description: Magnetic Resonance Images parallel simulator
- Deployment and status
 - MPI simulator implemented
 - Some performance study lead on local cluster
 - Tests on CINES supercomputers
- Users
 - Very large potential community
 - Today, only developers (5 users)
 - 1000 to 2000 jobs this year, minutes to weeks per job
- Plans: open the simulator as soon as gridification is achieved
- Problems no MPI-enabled resources available on EGEE infrastructure

- Name: xmipp_MLrefine
- Contact: Angel Merino, AJ.Merino@cnb.uam.es
- Description: Macromolecular 3D structure analysis
- Deployment and status
 - Recently ported to LCG2 and tested both on Clermont and Madrid clusters
- Users
 - Developers
 - One experiment corresponds to about 500 jobs and one week of computations on Madrid site
- Plans: To shorten experiments time by using more resources
- Problems: Resource shortage

- Name: gPTM3D
- Contact: Cécile Germain-Renaud, germain@lal.in2p3.fr
- Description: radiological data interactive segmentation and analysis
- Deployment and status
 - Application ported to LCG2 on top of the interactive job submission service
 - Deployed on Orsay resources
- Users
 - Developers
 - Potential medical users
- Problems: Interactivity made difficult due to bypass-based communication performance limitations



- **What EGEE offers external biomedical projects**

- **Access to large-scale infrastructure**

- Thousands of processors and 1/3 petabyte online data storage

- **Production ready grid middleware**

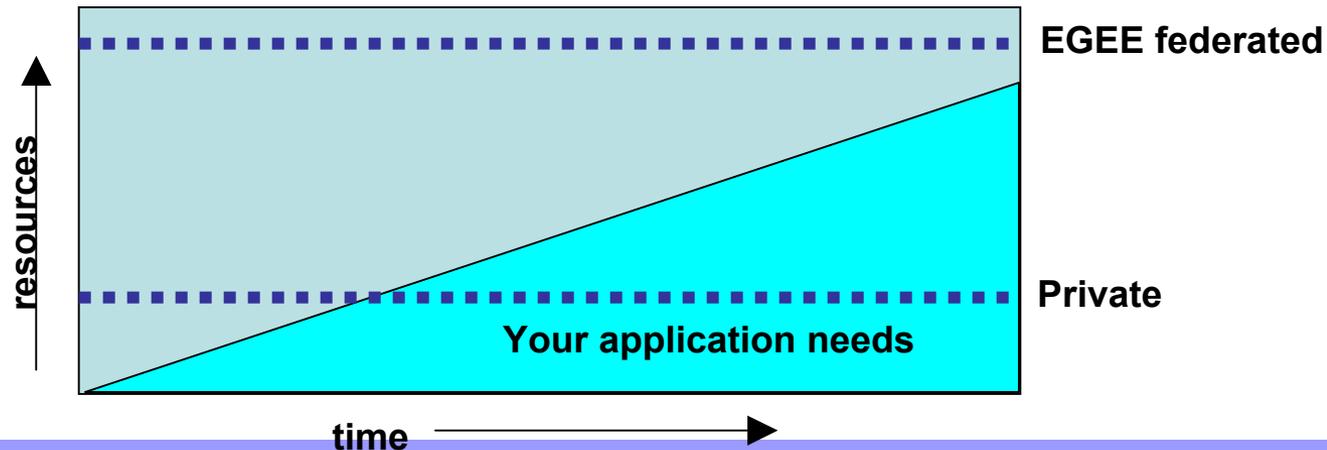
- More than 3 years of large-scale testing/deployment experience

- **Grid expertise**

- Small team of technically competent people ready to help applications get up and running

- **Training**

- **Applications in production usage have to contribute computing resources for production demands**



- European grid projects:
 - Mammogrid (breast cancer)
 - SIMDAT (drug discovery for neglected diseases)
- National grid projects
 - UK Mygrid (virtual laboratory)
 - French Rugby (bioinformatics services)
- Regional grid projects
 - Auvergne regional grid

Grids for rare diseases and diseases of the developing world

In silico drug discovery process
(EGEE, Swissgrid, ...)

An open-source shot in the arm?

June 10th 2004, The Economist print edition

www.economist.com/displaystory.cfm?story_id=2724420

Clermont-Ferrand

SCAI Fraunhofer

Swiss Biogrid consortium

Support to local centres in
plagued areas (genomics
research, clinical trials and
vector control)

Local research centres
in plagued areas

Interested to join ?

Contact V. Breton (breton@clermont.in2p3.fr)
or M. Hofmann (martin.hofmann@scai.fhg.de)

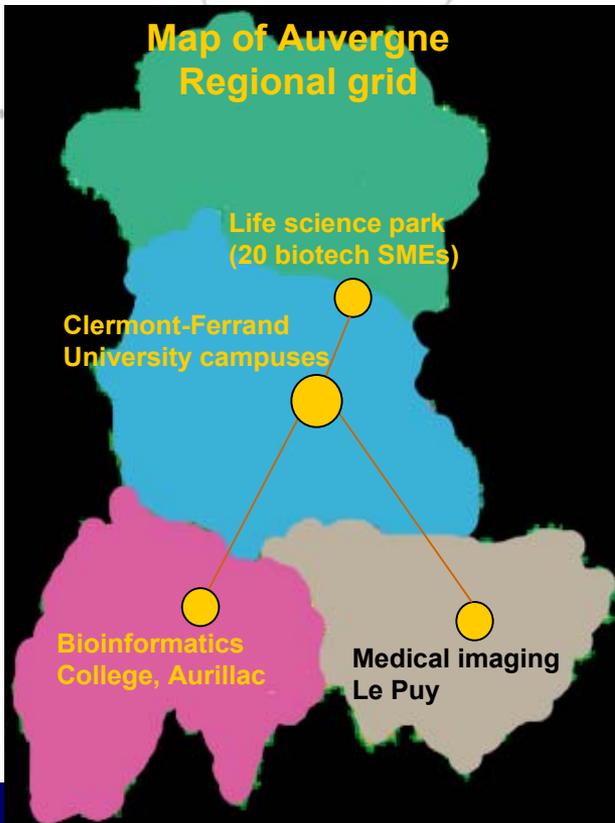
The grid impact :

- Computing and storage resources for genomics research and in silico drug discovery
- Cross-organizational collaboration space to progress research work
- Federation of patient databases for clinical trials and epidemiology in developing countries

- Mygrid is a virtual laboratory for data intensive biological analysis (www.mygrid.uk.org)
 - Pipelines, experiments as workflows
 - Adhoc exploratory investigative workflows
 - Low level of entry
 - Collection of components for assembly
 - Foundations for sharing knowledge and sharing experimental objects
 - Multiple stakeholders
- EGEE is an infrastructure providing resources and middleware on which to deploy virtual laboratories like Mygrid
-

INSTRUIRE : Regional Auvergne Infrastructure

- Operate at a regional level computing and storage resources to meet growing needs of
 - Research labs
 - Universities
 - Public administrations and services (hospitals,...)
 - Small and Medium Enterprises
- Favors regional-international collaborations



e.g. Federation of patient databases for clinical trials and epidemiology in developing countries (Relational DB, SRB)

Preparation and follow-up of medical missions in developing countries of the French NPO "Chain of Hope"

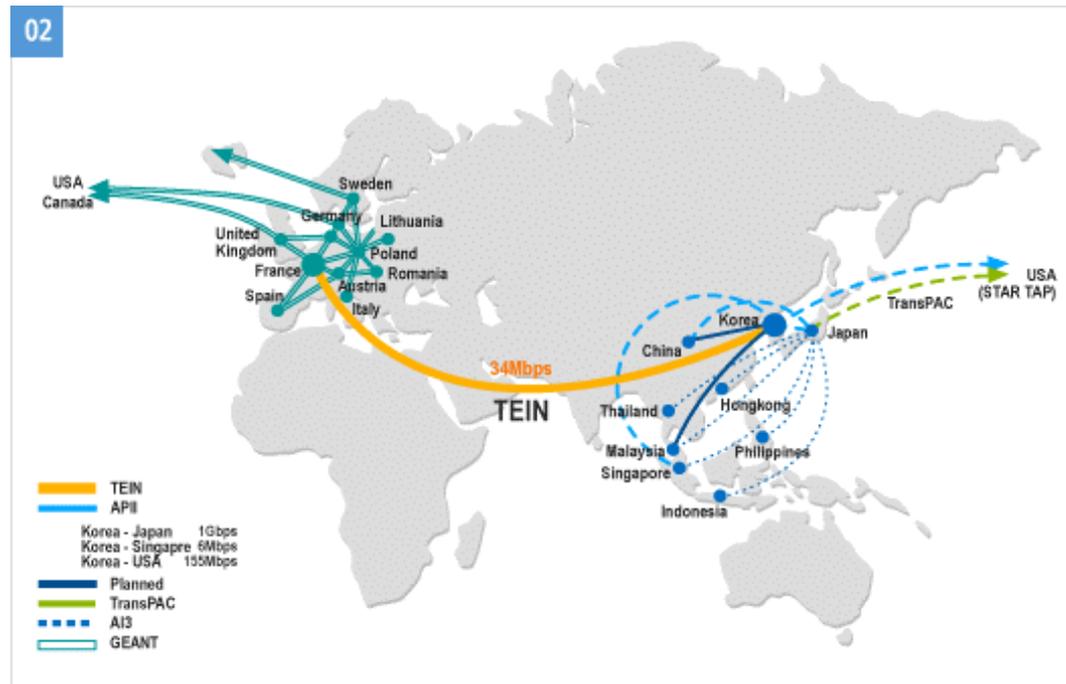
Support to local medical centres in terms of second diagnosis, patient follow-up and e-learning

current actions:

China: paediatric neurosurgery

Burkina Faso: ophthalmology

- **MoU** between EGEE and Chonnam National University-Kangnung National University-Sejong University Collaboration (CKSC)
- HEP applications:
 - development of the analysis system for ALICE experiment.
- Biomedical applications:
 - DNA and protein data analysis and Gene Regulation Bioinformatics.



HealthGrid initiative



- To provide a place of dialog and exchange between European and international projects
 - Web site: www.healthgrid.org
- To produce collaborative documents
 - HealthGrid White Paper ([download](#))
- To organize conferences and workshops on Health grids

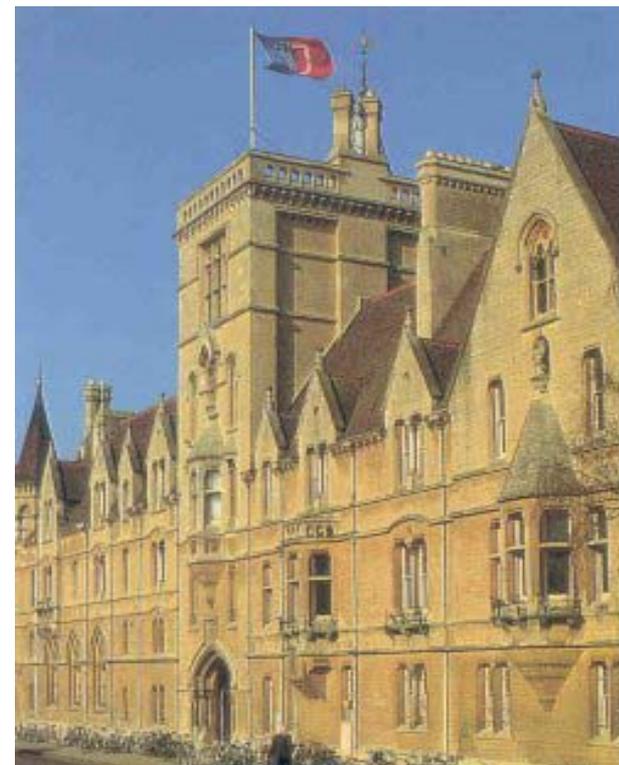
Third European HealthGrid Conference

Location : Oxford (UK)

Dates: April 7th – 9th 2005

Information and registration (available soon):

<http://oxford2005.healthgrid.org>



- EGEE project – www.eu-egee.org and its applications area (NA4)
- EU DataGrid – www.eu-edg.org
- The HEP LCG project www.cern.ch/lcg
- Questions to breton@clermont.in2p3.fr or project-po@cern.ch

- EGEE is expected to deliver a production Grid infrastructure. Life sciences are among the pilot applications to guide implementation and certify performances
- The project started 6 months ago
 - We have a running grid service based on LCG-2
 - All EGEE activities are well advanced
 - Next generation middleware being designed – first prototype made available to applications
- EGEE is interested to further explore possible new collaborations with international partners in life sciences
- Many thanks for your kind invitation!