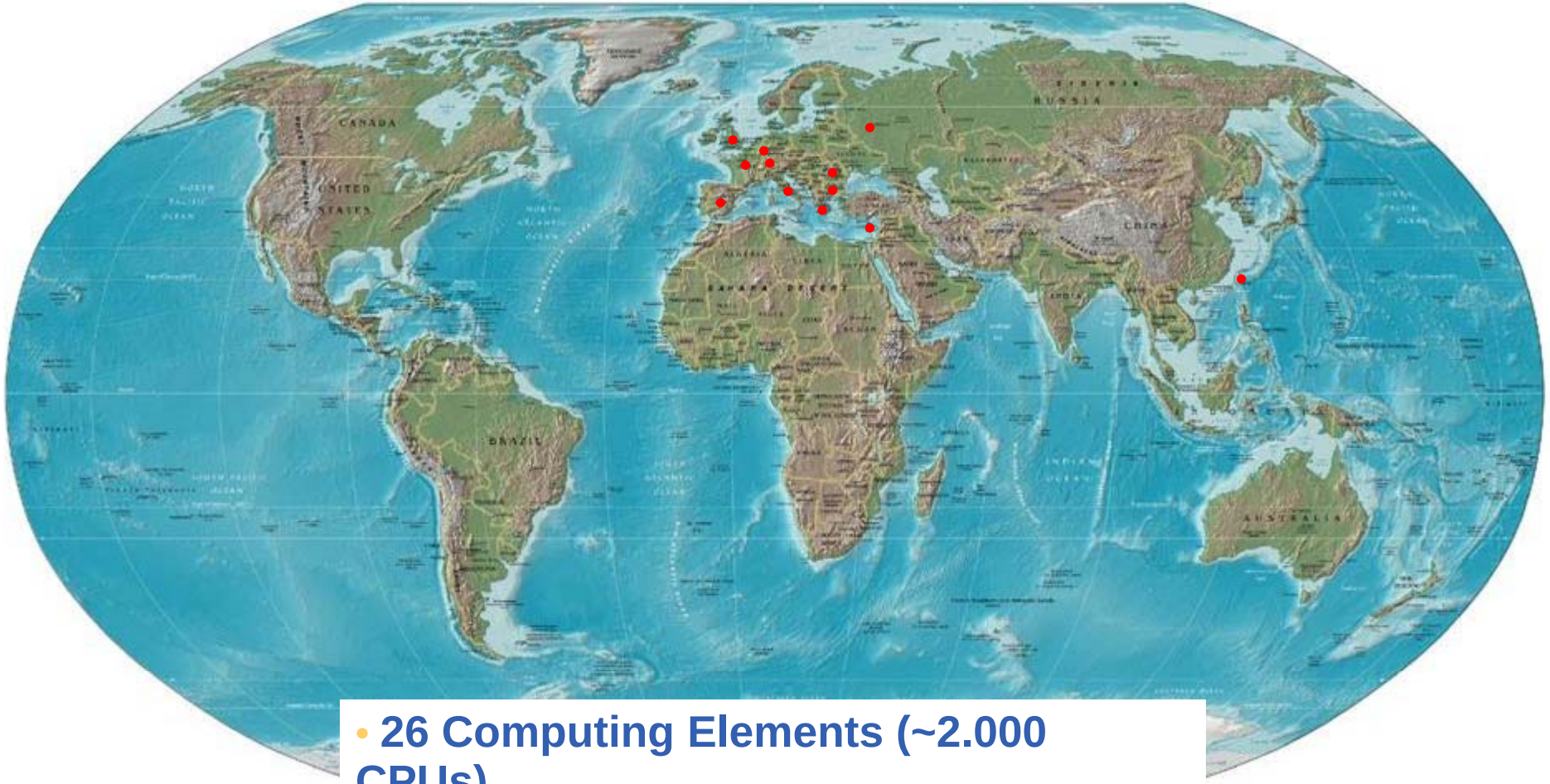
BioMedical Applications

Resources

Applications: deployment & success stories

The BioMed VO at a glance (1/2)

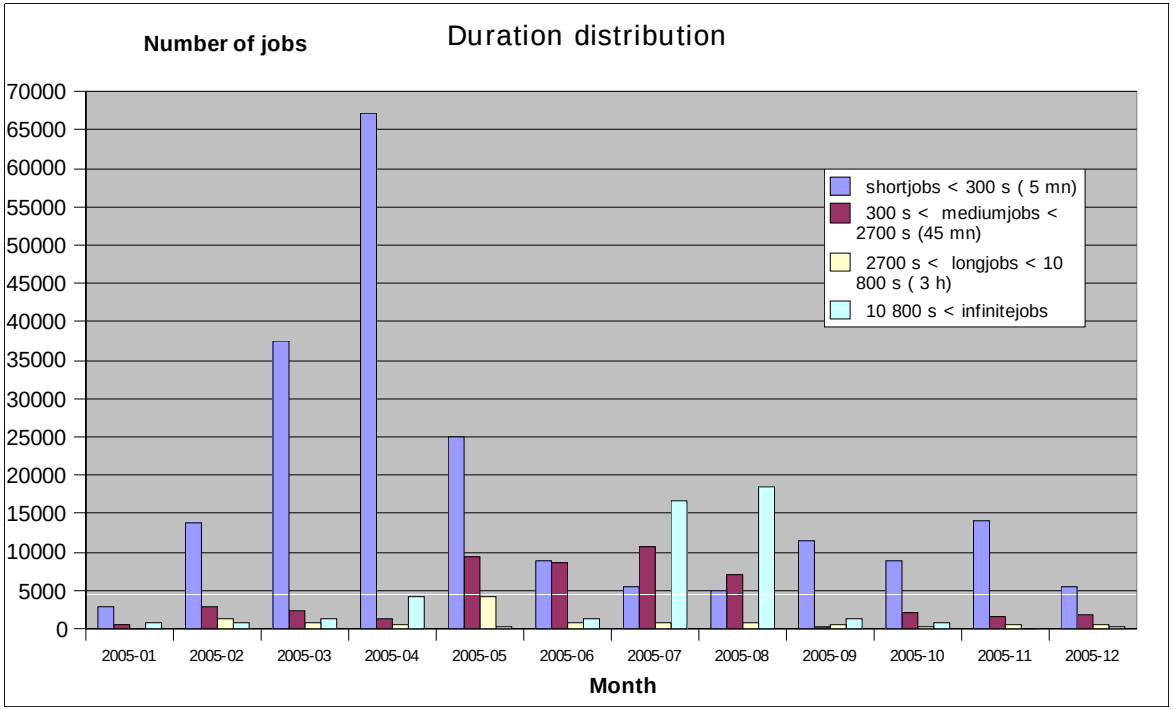
Enabling Grids for E-sciencE



- 26 Computing Elements (~2.000 CPUs)
- 28 Storage Elements (~21 TB disks)
- in 12 countries

The BioMed VO at a glance (2/2)

- 76 users
- 12 countries
- 18 research labs
- 12 Applications
- ~400,000 jobs launched since 04/2004
- ~100 CPU years



Pilot Applications

- **There are 3 pilot applications on the EGEE grid platform, running since the beginning of the project (April 2004)**

- **CDSS: Clinical Decision Support System**
  - Contact: Ignacio Blanquer (iblanque@dsic.upv.es)
- **GATE: Geant4 Application for Tomographic Emission**
  - Contact: Lydia Maigne (maigne@clermont.in2p3.fr)
- **GPS@: Bioinformatics grid portal**
  - Contact: Christophe Blanchet (christophe.blanchet@ibcp.fr)

Clinical Decision Support System

- **Scientific objectives**

  Extract clinically relevant knowledge from a large set of information with the objective of guiding the practitioners in their clinical practice. Similar (but non computer-based) systems exist since the 1950s.

  Example: "what are the genetic factors that *can be* involved in schizophrenia"?

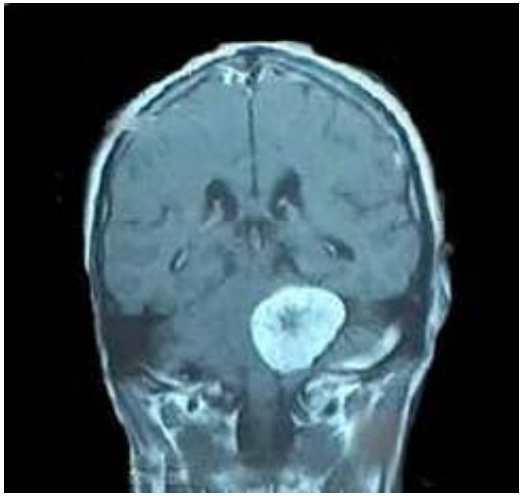
  CDSS does reinforce human decision by improving factors such as sensitivity, sensibility, and working conditions.

- **Method**

  Starting from **trained databases** such as:

  - classification of tumours soft tissues
  - classification of thalassemia and other anemia

  Use **classifier engines** and compare to annotated databases to classify data.



CDSS

- **Grid added value**
  - Ubiquitous access to distributed databases and classifier engines.
  - Use of grid information system at application level to publish and discover data sources and engines.
  - Automatic management of login and security.
- **Results and perspectives**
  - 12 classification engines available.
  - 1000 medical cases registered.
  - A web interface eases the access to the application.
  - These data and analysis tools are available for all users of the system (2 different user communities) independently of their actual location.
  - Dynamic discovery of all engines can be implemented on top of the grid information system.
  - Accounting will be provided by the grid.

GEANT4 Application to Tomography Emission

- **Scientific objectives**

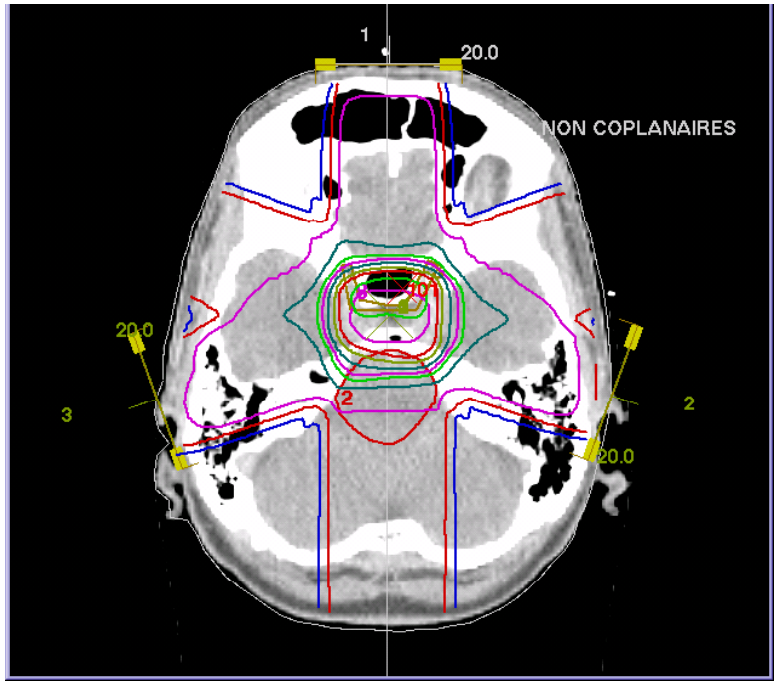
  Radiotherapy planning for improving the treatment of cancer by ionizing radiations of the tumours.

  Therapy planning is computed from pre-treatment MR scans by accurately locating tumours in 3D and computing radiation doses applied to the patients.

- **Method**

  GEANT4 base software to model physics of nuclear medicine.

  Use **Monte Carlo simulation** to improve accuracy of computations (as compared to the deterministic classical approach)



GATE

- **Grid added value**

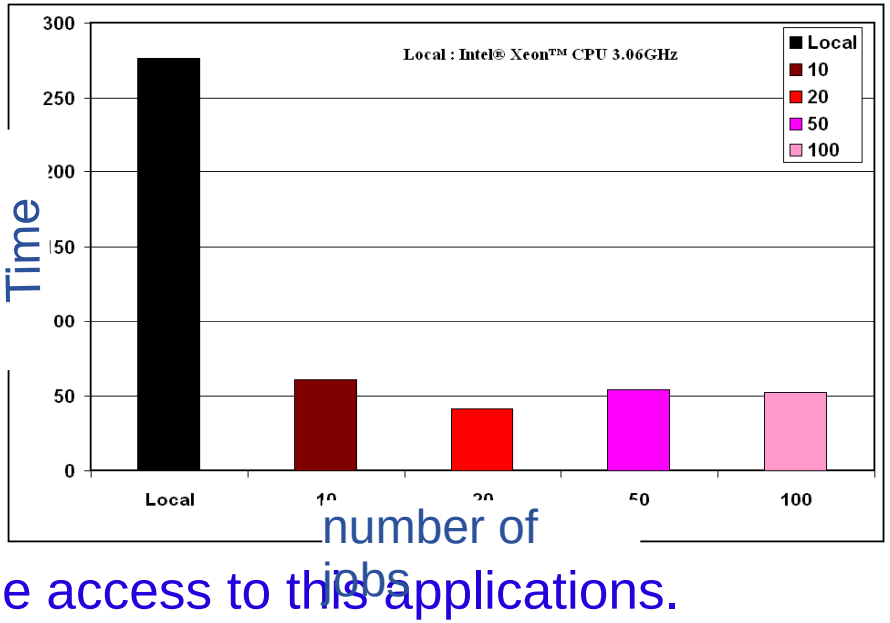
  Splitting the random number sequences needed for Monte Carlo simulations enables independent computations from different seeds.

  Different grid computing resources are used to perform the computations. This parallelization reduces the total computation time.

  Tread-off between level of splitting and number of jobs to process.

- **Results and perspectives**

  The computation time was reduced although not sufficiently for clinical practice: further optimisations are going on.



  A portal has been created to ease the access to this applications.

  A large community of users is interested in GATE.

Grid Protein Structure Analysis

- **Scientific objectives**

  Bioinformatic analysis of data produced by complete genome sequencing projects is one of the major challenge of the next years. Integrating up-to-date databanks and relevant algorithms is a clear requirement of such an analysis. Grid computing, such as the infrastructure provided by the EGEE European project, would be a viable solution to distribute data, algorithms, computing and storage resources for Genomics. Providing bioinformatician with a good interface to grid infrastructure will also be a challenge that should be successful. GPS@ web portal, **Grid Protein Sequence Analysis**, aims to be such an user-friendly interface for these grid genomic resources on the EGEE grid.

- **Method**

  A well-known web interface eases the access to the algorithms offered.

  Protein databases are stored on grid storage as flat files.

  Most protein sequence analysis tools are reference **legacy code** that is run unchanged. This tools are **wrapped in grid jobs** to be executed on grid resources.

  The algorithms output are analysed and displayed in graphic format through the web interface.

GPS@

- **Grid added value**

  The NPS@ portal records **3000 hits a day** and is limited in the size of the databanks and the kind of computations performed by local resources.

  The grid version, GPS@, can:

  - for **biological data**: provide Biologist with a convenient way to **distribute and to access to international databanks**, and to store **more and larger** of these databanks
  - for **bioinformatic algorithms**: allow **each portal user** to compute **larger datasets** with the available algorithms through **larger bioinformatic computations**
  - Open to a **wider user community**.

- **Results and perspectives**

  9 world-used bioinformatic softwares have currently been gridified: such as BLAST, CLUSTALW, PattInProt, …

  GPS@ is stressing the grid infrastructure with a large number of rather short jobs (few minutes each).

  Optimizations are worked on to:

  - Speed-up access to databases.
  - Lower short jobs latencies.
  - Processing data or software dependent jobs (workflow)

Internal Applications (1)

- **There are 9 internal applications on the EGEE grid platform, carried by the partners of the EGEE project**

- **SiMRI 3D: Magnetic Resonance Image simulator**
  - Contact: Hugues Benoit-Cattin
- **gPTM 3D: interactive radiological image visualization and processing tool**
  - Contact: Cécile Germain-Renaud
- **xmipp_MLrefine: Macromolecular 3D structure analysis**
  - Contact: Angel Merino
- **GridGRAMM: Molecular Docking web**
  - Contacts: Jose R Valverde, David Garcia
- **GROCK: Mass screenings of molecular interactions web**
  - Contacts: Jose R Valverde, David Garcia

Internal Applications (2)

- **Xmipp_assign_multiple_CTFs : Micrographia CTF calculation**
  - *Contact: Jose R Valverde*
- **Pharmacokinetics: Contrast agent diffusion in abdominal MR Images**
  - Contact: Ignacio Blanquer
- **Docking platform for tropical diseases: grid-enabled docking platform for in sillico drug discovery**
  - Contact: Nicolas Jacq
- **Bronze Standard: Evaluation of medical image registration algorithms**
  - Contact: Tristan Glatard

Pharmacokinetics

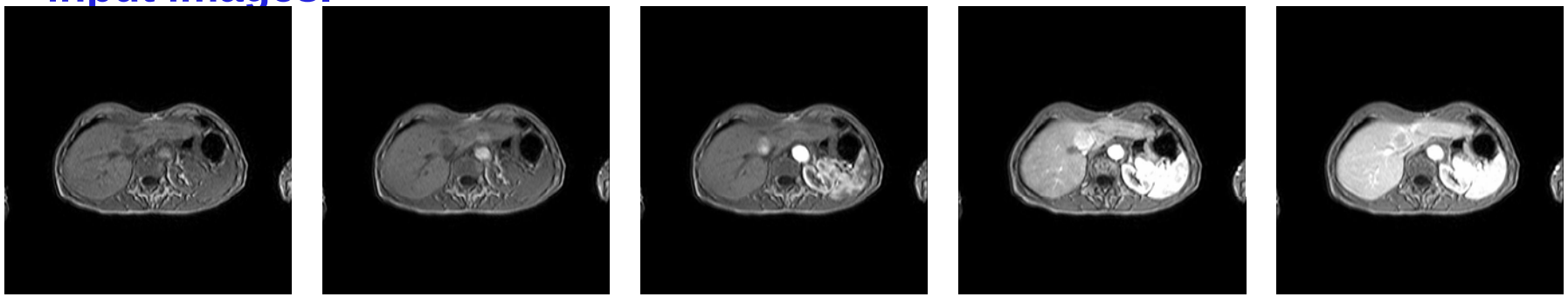
Co-registration of Medical Images

- **Scientific objectives**

  The study of **Contrast Agent Diffusion** can characterize tumour tissues without requiring biopsy. The process requires obtaining a sequence of MRI volumetric images. Before analyzing the variation of each voxel, **images must be co-registered** to minimize deformation due to different breath holds.

- **Method**

  The co-registration in the abdomen requires **deformable registration** methods. The Sequence of Images is co-registered with respect to the first volume. Voxel X,Y in all images must refer to the same body location. **Co-registration is compute intensive, especially when dealing with many input images.**



Pharmacokinetics

- **Grid added value**

  The grid is used for processing of **compute intensive** co-registration and generation of diffusion maps for the 3D MRI Studies.

  The grid resources are used **in parallel to perform independent computations** on different input data set.

- **Results and perspectives**

  Last clinical test: 12 patients with 13 MRI studies each. Each study comprises 24 512x512 12-bit slices.

  Processing of the registration algorithm in the last configuration takes around 1 hour per Volume (~12 hours per study).

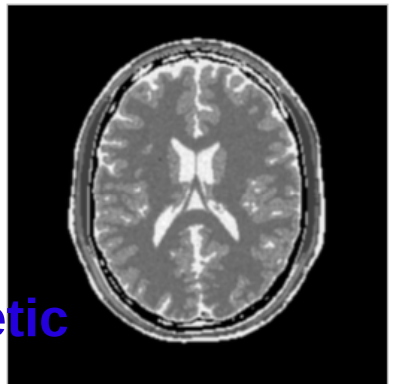
  The registration parameters (tolerance and number of iterations) were tuned with four possible combinations of values.

  Through an easy GUI, the data was Uploaded (registered) in the Grid, jobs were created, monitored, and results retrieved.

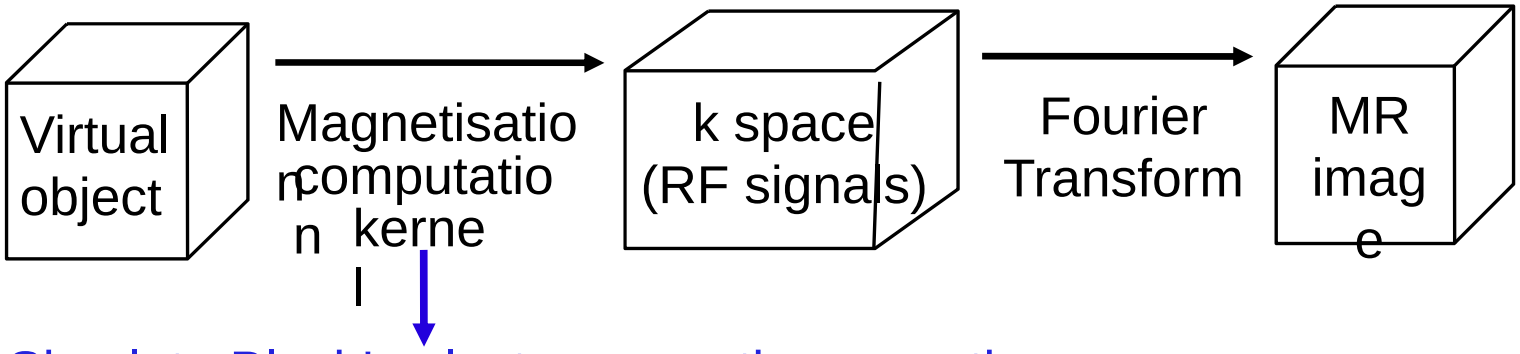
  One combination of parameter took 2 hours (**72 times faster than with a single computer**).

3D Magnetic Resonance Image Simulator

- **Scientific objectives**
  - **Better understand MR physics.**
  - **Study MR sequences in-silico.**
  - **Study MR artefacts.**
  - **Validate MR Image processing algorithms on synthetic yet realistic images.**



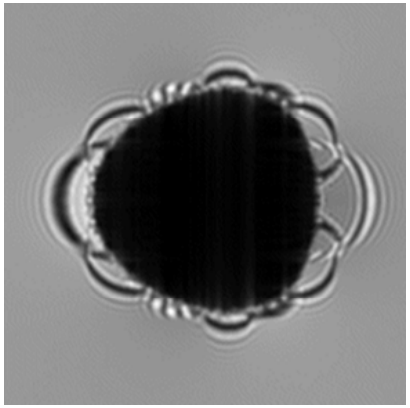
- **Method**



Simulate Bloch's electromagnetism equations:
parallel implementation to speed-up computations.

- **Grid added value**
  - Speeds up the simulation time
  - Enables simulation of high resolution images
  - Offers an access to MPI-enabled clusters
  - Offers a MRI simulation access to a wide variety of users
- **Results and perspectives**
  - Tractable computation time for medium size images:

	EGEE, IN2P3 (9 nodes)	**CINES** (64 nodes)	**PIV – 3 Ghz** 2 Go RAM
Image, 512^2	68 min	15 min	382 min
Volume, 64^3	68 min	13 min	382 min



  - Development of a portal to ease access to the application.
  - Implementation of new artefacts.

gPTM3D

3D Medical Image Analysis Software

- **Scientific objectives**

  Interactive volume reconstruction on large radiological data.

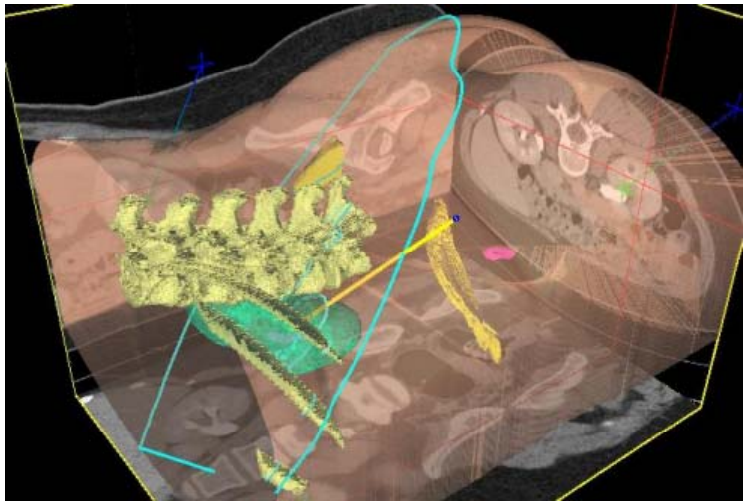
  PTM3D is an interactive tool for performing computer-assisted 3D segmentation and volume reconstruction and measurement (RSNA 2004)

  Reconstruction of complex organs (e.g. lung) or entire body from modern CT-scans is involved in augmented reality use case e.g. therapy planning.

- **Method**

  Starting from an hand-made rough Initialization,a **snake-based algorithm** segments each slice of a medical volume.

  3D reconstruction is achieved in **parallel** by triangulating contours from consecutive slices.



gPTM3D

- **Grid added value**

  Interactive reconstruction time: less than 2mins and scalable.

  Permanent availability of resources for fast reconstruction **and access** to any user at a non grid-enabled site (*e.g.* hospital).

  Close integration with a personal computer workflow.

  Unmodified medically optimized user interface.

- **Results and perspectives**

  The application was successfully ported and **demonstrated** at the first EGEE conference in Feb. 2005 and first EGEE review. The application is accessible to a wide community of **medical users**.

  Issues

  - Submission latency: scheduling agents, interest from other applications
  - Streams to/from non EGEE-enabled sites: specific protocol, Crossgrid glogin will be considered
  - Possible benchmark for Network QoS protocols (SA2)
  - Resource access QoS: ongoing work

xmipp_ML_refine

Macromolecules structure analysis from electron microscopy

- **Scientific objectives**

  Cryo-electron microscopy allows structural characterization of large biological assembly. Combining different views of a specimen enable **3D reconstruction** of **molecular structural information**.

  Macro molecules structure information is useful for studying molecules interactions and chemical properties of molecules.

- **Method**

  Multi-reference refinement of electron microscopy structures is achieved through a **maximum likelihood** statistical approach is used to find the most likely **model describing the experimental data**.

  This algorithm is less sensitive to false maxima of the correlation functions used in the classical approach.

xmipp_ML_refine

- **Grid added value**

  This application is very **compute intensive**: 2D analysis of multiple structures can take in the order of one to several weeks on a single CPU. 3D analysis is even more costly.

  On a grid, computation can be split in independent jobs that are executed in **parallel** on different **resources**.

- **Results and perspectives**

  First results on 2D analysis have shown significant time gains. For the same computation task, two months are needed on a local cluster (20 CPUs) versus half of it in the grid environment (one month).

  The algorithm is still being optimised and ported to the 3D case for further computations.

  Moreover an MPI implementation is currently being that should significantly improve the computation time.

Electron microscope images correction

- **Scientific objectives**

  Electron microscopy images are **impaired** by the electron sources and the defocus of magnetic lenses used in experimental practice.

  The image aberrations are described by a Contrast Transfer Function (**CTF**) that need to be estimated to fix images. The CTF is classically described by a parametric function.

  CTF estimation lead to drastic **image enhancement**.

- **Method**

  Auto regressive modelling is used to estimate parameters of the CTF and produce more reliable results than classical Fourier transform-based approaches.

xmipp_ML_CTFs

- **Grid added value**

  This application is very **compute intensive**: the functional is complex and costly to evaluate, and the optimisation process slow.

  A typical execution would take in the order of 2 months on a single CPU.

  Computations can be independently **parallelised** on different grid resources.

- **Results and perspectives**

  A typical execution is ran in about 2 months on a single CPU, 2 days on a local 20-CPUs cluster, and **14 hours on the grid**.

In silico Drug Discovery

- **Scientific objectives**

  Provide docking information helping in search for new drugs.

  Biological goal: propose new inhibitors (drug candidates) addressed to neglected diseases.

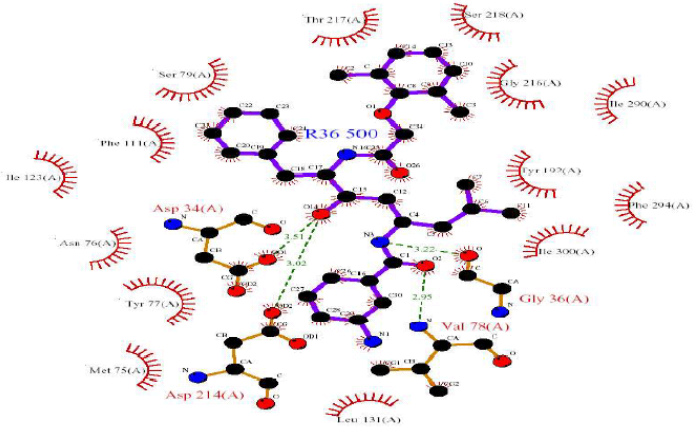
  Bioinformatics goal: *in silico* virtual screening of drug candidate DBs.

  Grid goal : demonstrate to the research communities active in the area of drug discovery the relevance of grid infrastructures through the deployment of a compute intensive application.

- **Method**

  Large scale molecular docking on malaria

  to compute million of potential drugs with some software and parameters settings. Docking is about computing the binding energy of a protein target to a library of potential drugs using a scoring algorithm.



In silico Drug Discovery

- **Grid added value**

  Drug discovery lead by pharmaceutical companies takes up to 12 years to complete. **Molecular docking** has the potential to drastically **speed-up** this process but considering large databases yield to **heavy computations**.

  The **computations** involved can be **distributed** on grid nodes by splitting the candidate drug input on different grid resources. The **data management** services will facilitate the storage and the post-processing of the output files

- **Results and perspectives**

  First experiments have shown that a limited size computation (10^5 candidate drugs tested against 1 protein target) are achievable in 2 days using the EGEE infrastructure compared to 6 months of CPU time involved.

  A full data challenge is planed that should involve $3x10^6$ candidate drugs to be tested against 5 protein target structures. The total computing time is expected to reach 80 years of CPU and 6 TB of storage.

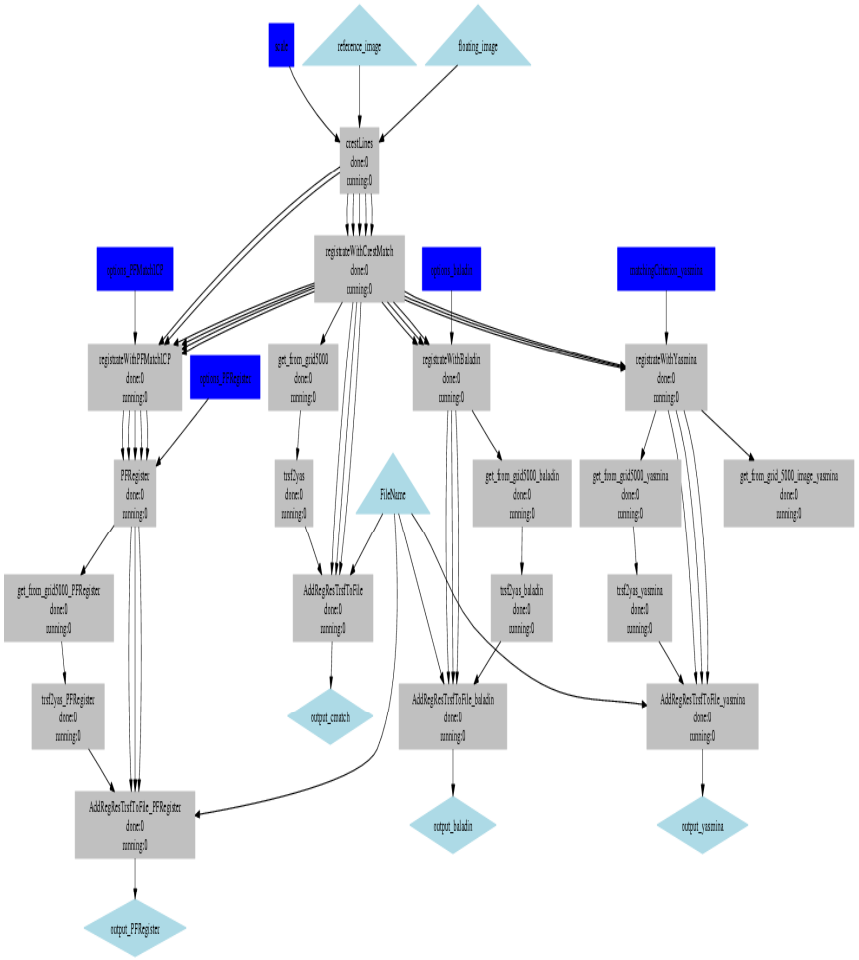
Bronze Standard

- ## Scientific objectives

  Evaluate medical image registration algorithms in the absence of reference gold standard.

- ## Method

  Compute a statistical **bronze standard** by exploiting the redundancies in transformations estimated using as many input pair of images to register and as many registration algorithms as possible. The application's complex workflow is handled by MOTEUR, a data-intensive workflow manager efficiently exploiting grid computing capabilities.



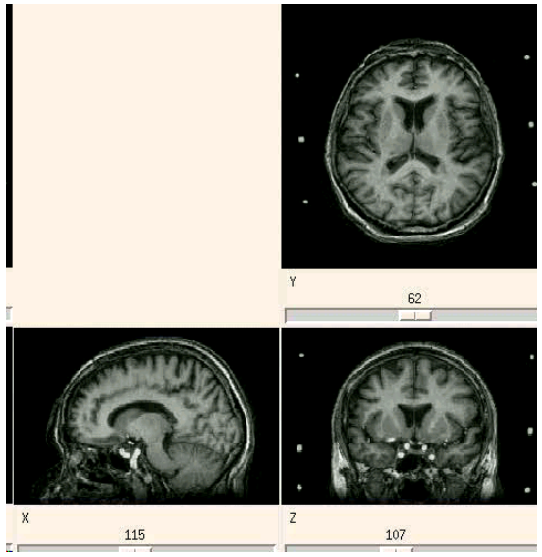
Bronze Standard

- **Grid added value**

  The more images can be used and the more registration algorithms can be found, the better the bronze standard will be. We are currently using hundreds of image pairs and 4 registration algorithms leading to thousands of registration computations. The grid is needed to handle the complex data flows of the application.

- **Results and perspectives**

  Through the MOTEUR workflow engine, the computation of a bronze standard takes a couple of hours only. It enables systematic assessment procedures for medical image registration algorithms. The application should be extended to integrate more algorithms. It should deploy an open portal to allow developers to integrate their own algorithm in the workflow.



External Applications

- **There are 9 external applications on the EGEE grid platform, carried by Research Institutes or Projects that are not partner of the initial EGEE project.**

- **SPLATCHE: genome evolution modeling**
  - Contacts: Nicolas Ray, Laurent Excoffier
- **The Mammogrid project**
- **MEX: motif extraction and protein classification**
  - Contact: Vered Kunik
- **bioDCV:**
  - Contact: Cesare Furlanello
- **Phylojava application: Phylogenetic analysis**
  - Contact: Manolo Gouy

Genome evolution modeling

http://cmpg.unibe.ch/software/splatche

- **Scientific objectives**

  Study **human evolutionary genetics** and answer questions such as the **geographic origin** of modern human populations, the **genetic signature** of expanding populations, the **genetic contacts** between modern humans and Neanderthals, and the expected null distributions of genetic statistics applied on **genome-wide data sets**.

- **Method**

  Simulate the past demography (growth and migrations) of human populations into a geographically realistic landscape, by taking into account the spatial and temporal heterogeneity of the environment.

  Generate the molecular diversity of several samples of genes drawn at any location of the current human's range, and compare it to the observed contemporary molecular diversity.

  SPLATCHE uses a region sampling Bayesian framework that requires 10^5 independent demographic and genetic simulations.

SPLATCHE

- **Grid added value**

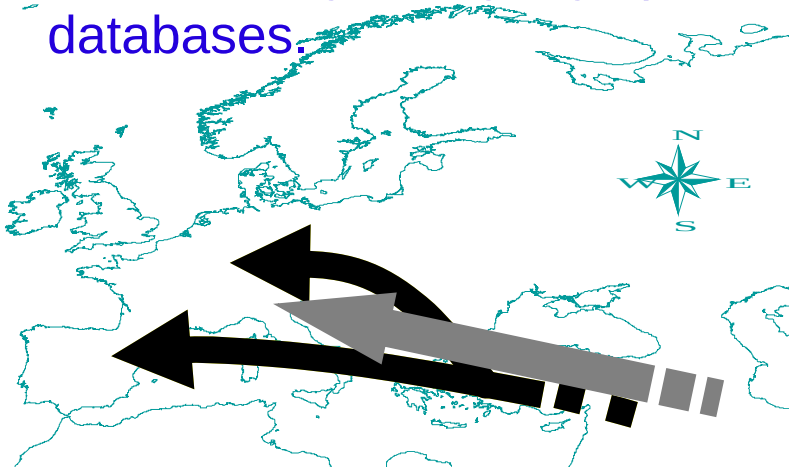
  Due to the Bayesian approach used, the SPLATCHE application is very **compute intensive**.

  On a grid the **independent simulations** can be executed in **parallel**.

- **Results and perspectives**

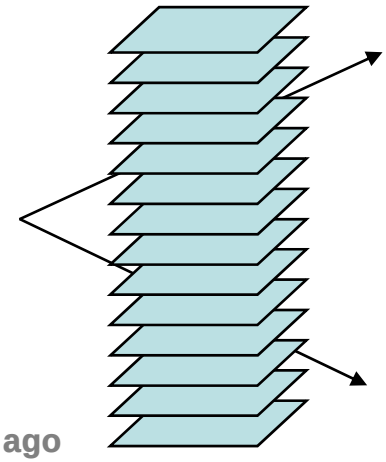
  Simulation of the two main colonization events in Europe (40k and 10k years ago).

  Plan to run global geographical dispersion simulation on enlarge databases.



Expansion of first farmers 8,000-3,000 years ago

Expansion of early modern Europeans 45,000-30,000 years ago



In black: relative proportion of genes brought by first farmers in various European regions (maximum=32%)

Probability distribution of demographic / genetic parameter of interest (admixture, migration, growth rates, etc…)

Mammogrid

- **Scientific objectives**

  To develop a European-wide database of mammograms that will be used to investigate a set of important healthcare applications as well as the potential of the Grid to support effective co-working between healthcare professionals throughout the EU.

  To deliver a medical information infrastructure that is service-based, grid-aware framework encompassing geographical regions of varying protocols, lifestyles and diagnostic procedures.

- **Method**

  The MammoGrid prototype uses the AliEn stack of the gLite middleware (as of 11.04) and a service-based database management system that is capable of managing federated mammogram databases distributed in Cambridge, Oxford, Udine and Geneva.

- **Grid added value**

  Storage elements for image files.

  Computing elements for imaging algorithms.

  Monitors for managing job submission and execution.

- **Results and perspectives**

Site	Number of Patients	Number of Image Files	Number of SMF Files	Associated Database Size	Storage Size
Cambridge	813	2798	2738	29 Mb	70 Gb
Udine	489	4663	2372	37 Mb	85 Gb

The average processing time for the core services are: (1) *Add* 8Mb DICOM file takes 7 seconds (2) *Retrieve* 8Mb DICOM file from a remote site takes 14 seconds (3) SMF workflow of *ExecuteAlgorithm* and *Add* takes 200 seconds. For querying:

Query	Cambridge	Udine	Num images	Num patients
By Id: Cambridge patient	2.654 sec	2.563 sec	8	1
By Id: Udine patient	2.844	3.225	16	1
All female	103	91	12571	1510
Age [50,60] and ImageLaterality=L	19.489	22.673	1764	357