LIFE SCIENCE APPLICATIONS

The life sciences are a major application area for the Enabling Grids for E-sciencE (EGEE) project. More than 30 separate applications are deployed or being ported. These applications place high demands on the middleware with strict security requirements, intensive data management and the execution of large numbers of small jobs.

The life science applications are established regular users of EGEE's infrastructure. The biomedical Virtual Organization is the second biggest consumer of infrastructure resources, after the four LHC experiments. The life science applications on the EGEE infrastructure are subdivided into three areas: **medical image processing**, **bioinformatics**, and **drug discovery**.

The **medical imaging** sector targets the computerised analysis of digital medical images. It includes medical data federation, compute-intensive medical procedures, processing large data sets and statistical studies over large populations. Specific applications include:

- **GATE** is a Monte Carlo-based simulator for planning radiotherapy treatments based on patient images. It uses the EGEE grid infrastructure to reduce the time needed to complete Monte Carlo simulations relevant for clinical use.
- The Clinical Decision Support System (**CDSS**) uses image classification based on expert knowledge to aid clinical decisions. The grid is exploited both for collecting large data sets and for efficiently training classification software over these large data sets.
- The **Pharmacokinetics** application studies the diffusion of a contrast agent in the liver from sequences of magnetic resonance images. Artefacts due to the movement of the patient make images directly incomparable. However the parallelised image co-registration computations running on the grid allow analysis of the sequence in a reasonable time.
- **SiMRI3D** is a Magnetic Resonance Imaging simulation for the production of artificial but realistic 3D Magnetic Resonance (MR) images to analyse images from perfectly known sources, study artefacts, and further develop and optimise MR sequences.
- The **gPTM3D** application allows interactive reconstruction of 3D medical images, e.g. for the volume reconstruction of large or complex organs. The quality-of-service required for interactivity means that some sites on the grid have to define a high-priority for this class of jobs.
- The **Bronze Standard** is an application to evaluate medical image registration algorithms. The amount of data to manipulate and the cost of computations are out of reach for standard computers, yet the application can easily be distributed over a grid.
- The **SPM** software package is used by the neurological research community for the early diagnosis of Alzheimer's disease. It is based on the comparison of the candidate case to a large set of normal cases. Grid technologies allow easy access to distributed data as well as to distributed computational resources.
- **SEE++** is a software program for the biomechanical 3D simulation of the human eye and its muscles. It simulates the common eye muscle surgery techniques in a graphic interactive way that is familiar to an experienced surgeon to deal with the support of diagnosis and treatment of strabismus.
- **ThIS** is a Therapeutic Irradiation Simulator based on the GEANT4 toolkit. It simulates the irradiations of living tissues with photons, protons or light ions beams for cancer therapy. Monte Carlo simulation is parallelized over grid resources for efficiency.

Application domain services are also developed to help the migration of application to the grid infrastructure:

- The Medical Data Manager (**MDM**) is a high level middleware service, tightly coupled to the gLite middleware, for secured medical data management. It proposes a DICOM-to-grid data management system interface, medical metadata management and high security.

- **MOTEUR** is a data intensive, gLite interfaced, workflow manager that is well suited to describe and enact image analysis pipelines on the grid. It produces a high level interface to the grid for end users and transparently exploits application parallelism to optimize performance.

EGEE III is co-funded by the European Commission under contract number INFSO-RI-222667

The **bioinformatics** domain studies genes, proteins, and other components of living organisms. These include studies of systems biology on the grid, molecular level oncology, genome wide association studies of diseases, protein and DNA binding in the cell nucleus and complete genome comparisons. Some examples of grid-enabled applications are:

- **Anchoring the nucleosome**: nucleosome formation and movement on DNA, of great biological significance, is one of the least well understood aspects of gene control. The last experiment was about three runs of 142,336 jobs each, for a total 24 years of computation.
- **Metagenomics**: one of the experiments is related to the analysis of the metagenome of the intestinal microbiota of 13 patients and controls, involving up to 300,000 sequences.
- **Grid Protein Sequence Analysis** is a bioinformatic Web portal providing a user-friendly interface to biological grid resources. Tools like BLAST, FastA, ClustalW, and databanks like Swiss-Prot or TrEMBL are available and integrated on grid through this portal.
- Grid methods for stability analysis of biomarkers: **BioDCV** is in a production version for predictive profiling on high-throughput technologies with computational procedures for complete validation for control of Selection Bias/Overfitting, stability analysis of Biomarker Lists.
- **Systems biology** investigates the behaviour of complex systems of interacting components. This application has explored the feasibility of simulating the behaviour of a large system (530,000 jobs, 1,000,000 results files, 102 years cumulated computation time), involving multiple signalling pathways relevant for human cancer, and can be used as a template for large-scale simulation of biological systems.
- **Genome-Wide Association Studies** of human complex diseases: for the first genome-wide haplotypes analysis on coronary artery disease, a sliding-windows haplotype approach has been applied to a case-control study of 1988 cases and 3004 controls genotyped for 377,857 Single Nucleotide Polymorphisms (SNPs).
- Large-scale **genome comparison** is dealing with the automatic clusterization and annotation of complete protein databases. The experiment is an “all-against-all” comparison of all protein sequences of 599 genomes.

The drug discovery sector aims to help speed up the process of finding new drugs through in silico simulations of protein structures and dynamics.

- The **WISDOM** initiative runs large-scale computations for the in silico drug discovery against emerging and neglected diseases. These virtual screening calculations determine how well certain drugs attach to specific sites on the target virus – those which dock are more likely to be active against the virus. It has been successfully deployed against malaria and avian influenza, with exciting results confirmed in vitro.
- **GridGRAMM** is a simple interface to do molecular docking on the Web. Results include a quality score and various access methods to the 3D structure of the complex. Molecular docking can be used for the study of molecular interactions, to analyze enzyme-substrate interactions, for drug design and to understand morbid mutant behaviour.
- The goal of **GROCK** (Grid Dock) is to provide an easy way to conduct mass screenings of molecular interactions using the Web by allowing users to screen one molecule against a whole database of known structures.

Application webpages

EGEE is keen to consider other applications. For further information on how to participate see **http://technical.eu-egee.org/index.php?id=392**.
More information about the applications running on EGEE be found on the EGEE website at **http://technical.eu-egee.org/index.php?id=148**.

Group contacts

Drug Discovery	**Vincent Breton (IN2P3-LPC),** email: breton@clermont.in2p3.fr
Medical Image Processing	**Johan Montagnat (I3S),** email: johan@i3s.unice.fr
Bioinformatics	**Christophe Blanchet (IBCP),** email: christophe.blanchet@ibcp.fr

EGEE III is co-funded by the European Commission under contract number INFSO-RI-222667