Medical Imaging Production and Services on the EGEE Grid

Johan Montagnat

EGEE Industry Day

Bratislava, September 19, 2007
History

- **DataGrid project (2001-2003): WP10 “biomedical applications”**
  - Grid infrastructures discovery
  - From “biology” to “biomedical” (life-science area)
  - First requirements analysis
  - Porting of first pilots and infrastructure testing

  - Entering production mode
  - Middleware steering and infrastructure testing role
  - More applications porting (12 applications in the biomed VO)

- **EGEE-II (2006-2008): NA4/biomed divided in 3 sectors**
  - Medical imaging, bioinformatics, drug discovery
The BioMed VO today

- > 90 Computing Elements (> 8 000 CPUs)
- ~21 TB disk space in Storage Elements
- in ~20 countries
• Regular infrastructure usage for production
  – > 100 users
  – > 12 applications
  – ~25 Kjobs per month
  – Generally not so compute intensive: many small jobs

• Much more than computing power expected
  – High level data management required (mandatory!)
  – Data protection
  – Sharing data, procedures and resources
  – Easing access to application code for non-specialists
  – ...

Biomedical applications in EGEE

Enabling Grids for E-sciencE
Medical imaging sector

- CDSS
  - UPV (University Polytechnic of Valencia)
- Pharmacokinetics
  - UPV (University Polytechnic of Valencia)
- GATE
  - LPC (CNRS – Clermont Ferrand)
- SiMRI3D
  - CREATIS (CNRS - Lyon)
- gPTM3D
  - LRI-LAL-LIMSI (CNRS - Orsay)
- Bronze Standard
  - I3S (CNRS – Sophia), INRIA
CDSS: 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 or classification of thalassemia and other anemia. Use classifier engines and compare to annotated databases to classify data.
• **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.
• **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)
SiMRI3D: MRI 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.
  - Paralle (MPI)l implementation to speed-up computations.
gPTM3D: radiology analysis

• **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.
Bronze Standard: assessment

- **Scientific objectives**
  Evaluate medical image registration algorithms in the absence of reference gold standard.

- **Method**
  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.
Application specific requirements

• “Embarassingly” parallel problem
  – WISDOM, GROCK, GATE, Pharmacokinetics, SPLATCHE, xmipp_ML...

• Fast turn over of short jobs
  – CDSS, GPS@

• Data confidentiality
  – GATE, gPTM3D, BronzeStandard ... *medical imaging*

• Interactivity
  – gPTM3D
Application specific requirements

• Fine grain parallelism (MPI)
  – SiMRI3D, xmipp_ML

• Workflow-based
  – BronzeStandard

• Portal
  – GPS@, SiMRI3D, BronzeStandard...
Application level services

- **Medical Data Manager**
  - Grid storage interface to DICOM databases

- **Data protection**
  - Encryption (storage/network protection) + access control (privacy protection). Two existing systems:
    - Hydra key store integrated in gLiteIO (JRA1)
    - Perroquet (Parrot based, IBCP)

- **Data parallel workflows**
  - Data intensive service composition-based applications

- **Fast turn over**
  - Short deadline jobs
Objectives

- Expose a **standard grid interface (SRM)** for **medical image servers (DICOM)**
- Use native DICOM storage format
- Fulfill medical applications security requirements
- Do not interfere with clinical practice
Interfacing sensitive medical data

- **Privacy**
  - **Fireman** provides file level **ACLs**
  - **gLiteIO** provides **transparent** access control
  - **AMGA** provides metadata **secured communication** and **ACLs**
  - **SRM-DICOM** provides on-the-fly **data anonymization**
    - It is based on the dCache implementation (SRM v1.1)

- **Data protection**
  - **Hydra** provides encryption/decryption **transparently**
Medical Data Registration

1. Image is acquired

2. Image is stored in DICOM server

3. glite-eds-put

3a. Image is registered

3b. Image key is produced and registered

4. Image metadata are registered

AMGA Metadata

DICOM server

gLiteIO server

Fireman File Catalog

Hydra Key store

Medical Imaging, Industry Day, Bratislava, September 19, 2007
Medical Data Retrieval

EGEE-II INFSO-RI-031688
Medical Imaging, Industry Day, Bratislava, September 19, 2007

1. get GUID from metadata
2. glite-eds-get
3. get SURL from GUID
4. request file
5. get file key
6. on-the-fly encryption and anonymization
7. get file key and decrypt

In-memory decryption

User Interface
Worker Node

Key store

gLiteIO client

gLiteIO server

Metadata DN-based ACL
File DN-based ACL

AMGA Metadata

Fireman

File Catalog

SRM-DICOM interface

Anonymization & encryption

DICOM server

Hydra

Key store
Enabling Grids for E-sciencE

~100 image pairs
~800 EGEE jobs

Image analysis pipeline

Params

A

PFRegister

GetFromEGEE

FormatConv

WriteResults

CrestLines

Service

B

Yasmina

GetFromEGEE

FormatConv

WriteResults

Baladin

GetFromEGEE

FormatConv

WriteResults

MultiTransfoTest

Accuracy Translation

Accuracy Rotation
Workflow management

- MOTEUR, I3S laboratory, CNRS
  - http://egee1.unice.fr/MOTEUR
- Service-based approach
  - Legacy code service wrapper
- Scufl language (myGrid / Taverna)
  - Pure data flow approach
- Grid submission interfaces
  - EGEE (LCG2, gLite)
  - Grid5000 (OAR, DIET)
- Transparent parallelism exploitation
  - Code and data parallelism
A workflow naturally provides application parallelization

MOTEUR transparently exploits 3 kinds of parallelism

- **Workflow parallelism**
  - $D_0$, $D_1$

- **Data parallelism**
  - $D_0$, $D_1$

- **Service parallelism**
  - $D_0$, $D_1$

Jobs grouping strategy in sequential branches in order to reduce grid latency
Performance results

- On the EGEE infrastructure (biomed VO)

![Graph showing performance results]

- JG = Job Grouping
- DP = Data Parallelism
- SP = Service Parallelism (pipelining)

Graph details:

- Execution time (hours) vs. number of images
- Lines represent different combinations of techniques:
  - NOP
  - JG
  - DP
  - SP+DP
  - SP+DP+JG

Legend:

- JG = Job Grouping
- DP = Data Parallelism
- SP = Service Parallelism (pipelining)
Short Deadline Jobs

• **Guarantee low-latency execution**
  - Submit-or-reject: Jobs execute or are refused immediately.
  - Complementary to job prioritization.

• **Torque + MAUI batch scheduler**
  - Used on most EGEE sites
  - Specific configuration (virtual processors allocation)
  - No interference with normal batch scheduling (shared proc. time)
  - Enables efficient processing of short tasks on grid

• **Special submission queues**
  - Three biomed sites support SDJ queues
  - Tight limits on CPU and wall-clock time
  - Access to queue also limited
Perspectives

• **More data protection**
  – Role-based access control
  – Considering multiple actors involved in clinical procedures
    (patient, doctor, surgeon, radiologist, nurse, anesthesist...)

• **Data semantics**
  – Metadata associated to data
  – Ontologies
  – Content-based data identification, indexing, mining...

• **User interfaces**
  – Integrated in end-users environments
  – Often specific to application

• **Touch the real end-users (physicians, biologists...)**