GGF 10

Biomedical applications on Grids
The EU Datagrid / Medigrid experience

Johan Montagnat

EU DataGrid work package on biomedical applications deputy manager

Medigrid manager

The EU DataGrid IST project

January 2001 – February 2004

- 5 Middleware workpackages
  - Jobs management
  - Data management
  - Information system
  - Fabric management
  - Mass storage
- 1 Deployment workpackage
- 1 Networking workpackage
- 3 Application workpackages
  - High energy physics
  - Earth observation
  - Biomedical applications = bioinformatics + medical imaging

WP10: Biomedical Applications
Objectives

- To demonstrate the relevance of grids for life science
  - Identify the need for grid technologies
  - Biomedical applications requirement collection
- To test the EDG middleware and feedback requirements to the middleware developers
  - Application deployment on the EDG testbed
  - Testbed 1 in year 2001 (Globus Toolkit 2)
  - Testbed 2 from September 2002 (Resource Broker, Grid Data Mirroring...)
  - Testbed 3 from October 2003 (Bug fixes and new functionnalities)
- To raise awareness on the impact of grids in the life science community
  - Dissemination: HealthGrid initiative, projects, etc
  - Result reports

Middleware capabilities

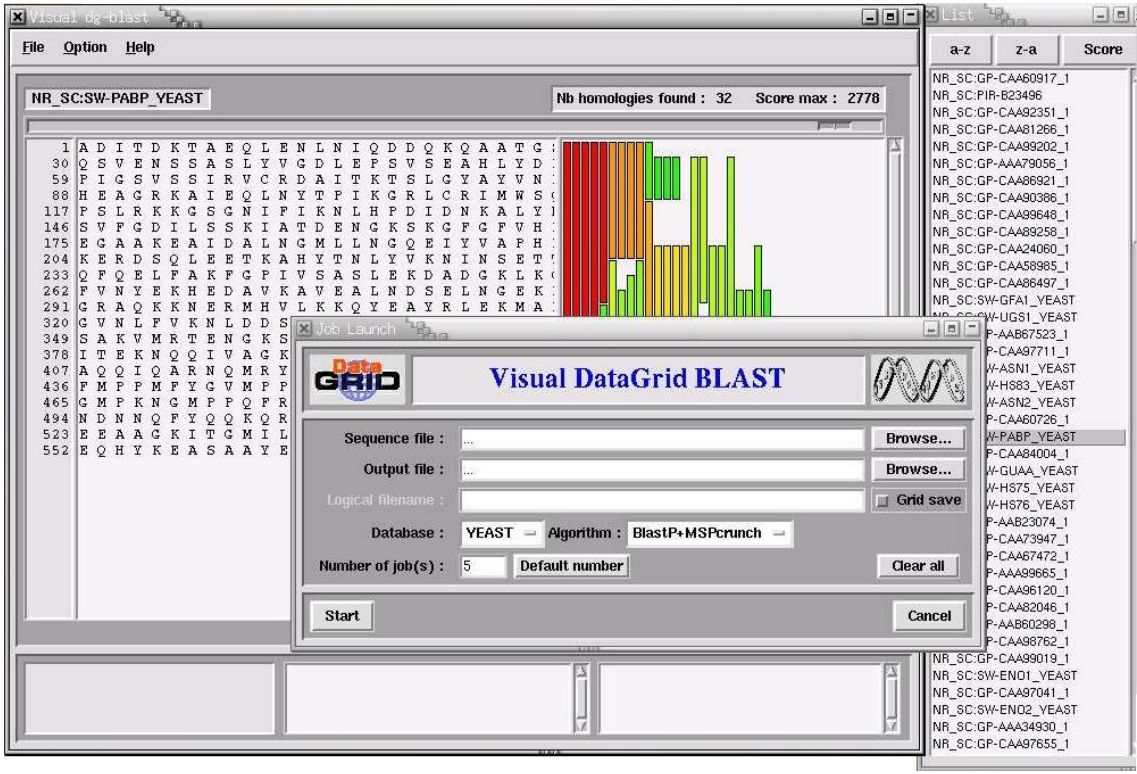
- Based on the Globus Toolkit 2 + Condor
- Job management
  - Resource broker, batch job submission
  - MPI jobs submission
  - Interactive jobs submission (shell based interface with stdin/out redirection)
- Data management
  - Replica Location Service (read-only replicas)
  - Metadata management
- Information system
  - RGMA: Relational Grid Monitoring Architecture
  - Code instrumentation for monitoring
- Fabric management
- Virtual Organisation Management System (Role-based security)

Biomedical application requirements

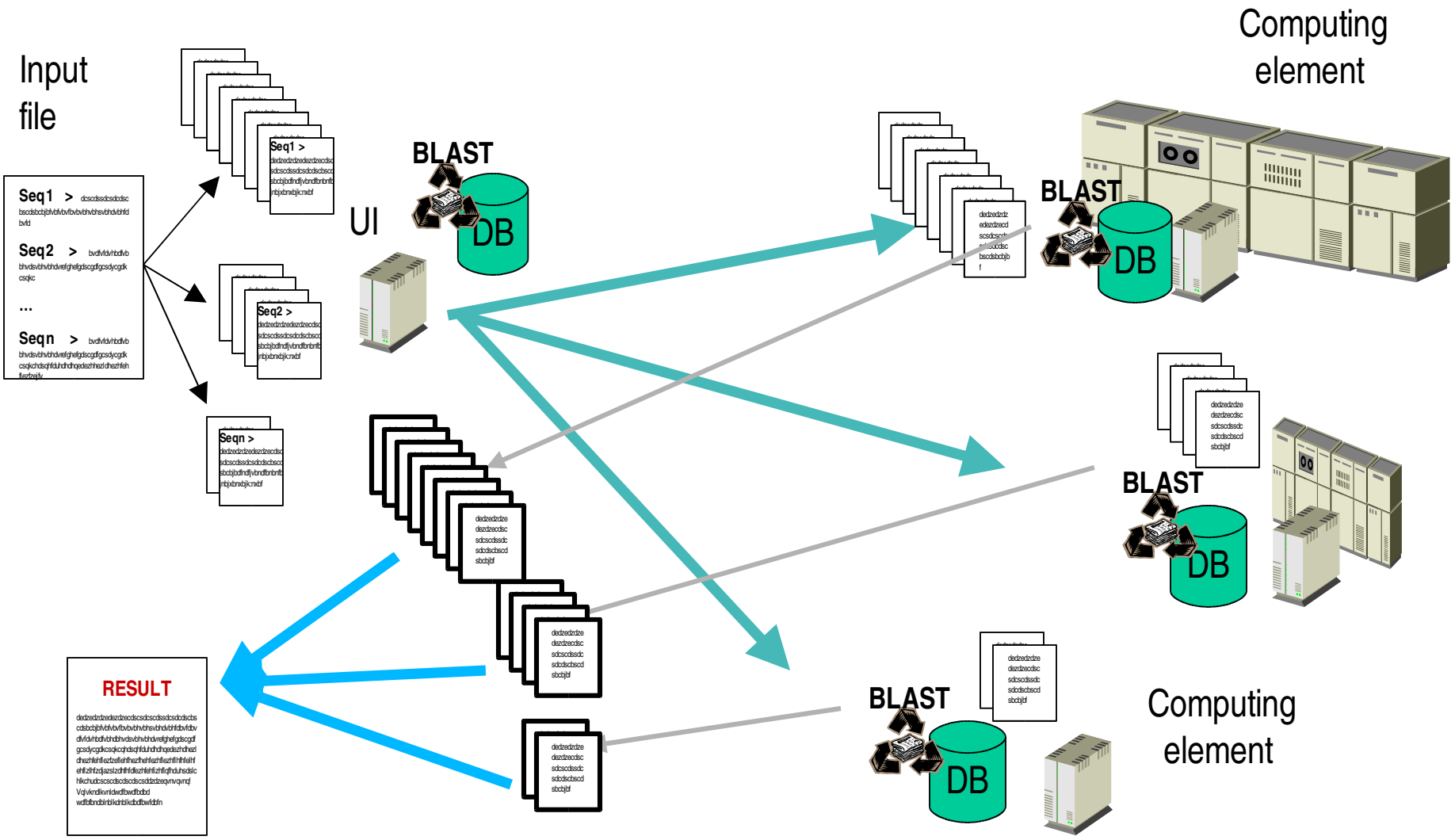
- Complex data requirements
  - Heterogeneous data formats (genomics, proteomics, image formats)
  - Frequent data updates
  - Complex data sets (medical records)
  - Security/privacy constraints
  - Long term archiving requirements
- Complex processing requirements
  - Bioinformatics: gene/proteome databases distributions
  - Medical applications (screening, epidemiology...): image databases distribution
  - Parallel algorithms for medical image processing, simulation, etc
  - Interactive application (human supervision or simulation)
  - Security/privacy constraints

BLAST: Bioinformatics on the EDG testbed

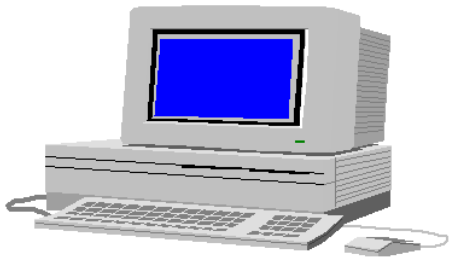
- BLAST is the first step for analysing new sequences: to compare DNA or protein sequences to other ones stored in personal or public databases.
- BLAST is costly and a good candidate for gridification:
  - Requires equipment to store databases and run algorithms
  - Requires manpower for system & network maintenance and frequent update of databases
  - Large user community

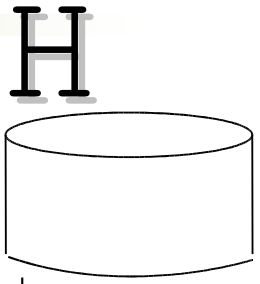


BLAST gridification



Medical image content-based queries





LFN	image	patient	hospital	...

Metadata

Medical images

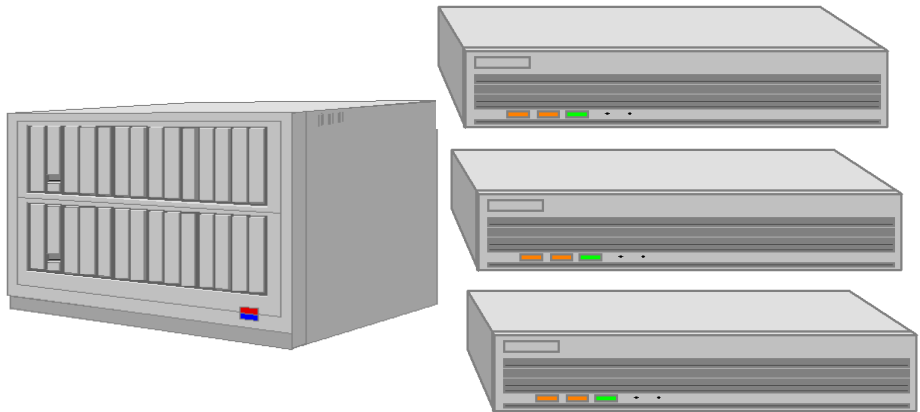
1. query

2. visualisation

5. best results visualisation

3. similarity search

4. scores

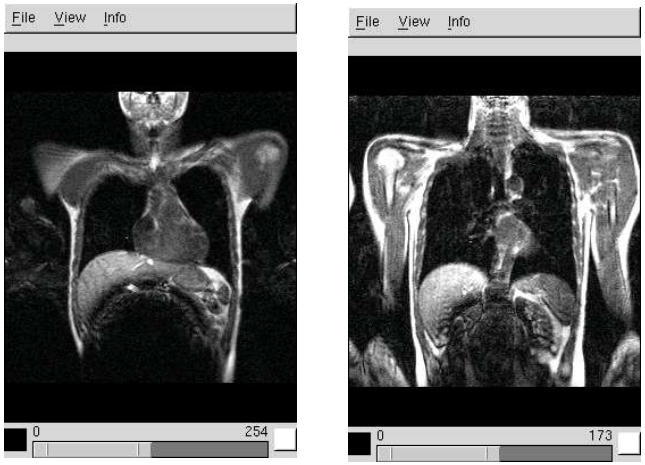


Medical image content-based queries

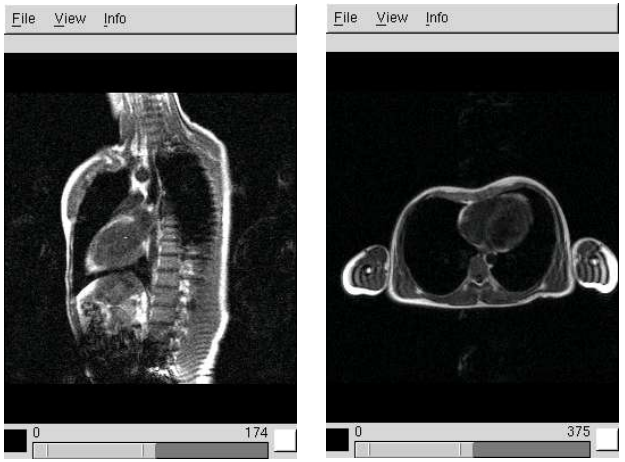
- Coefficient of correlation $$\eta^2(I|J) = 1 - \frac{1}{\sigma_I^2} \sum_j p_j \sigma_{I|j}^2$$
- Mutual information $$H(I|J) = -\sum_i \sum_j p_{ij} \frac{p_{ij}}{p_j}$$

Results: running tens of image similarity measurement in parallel

Source image



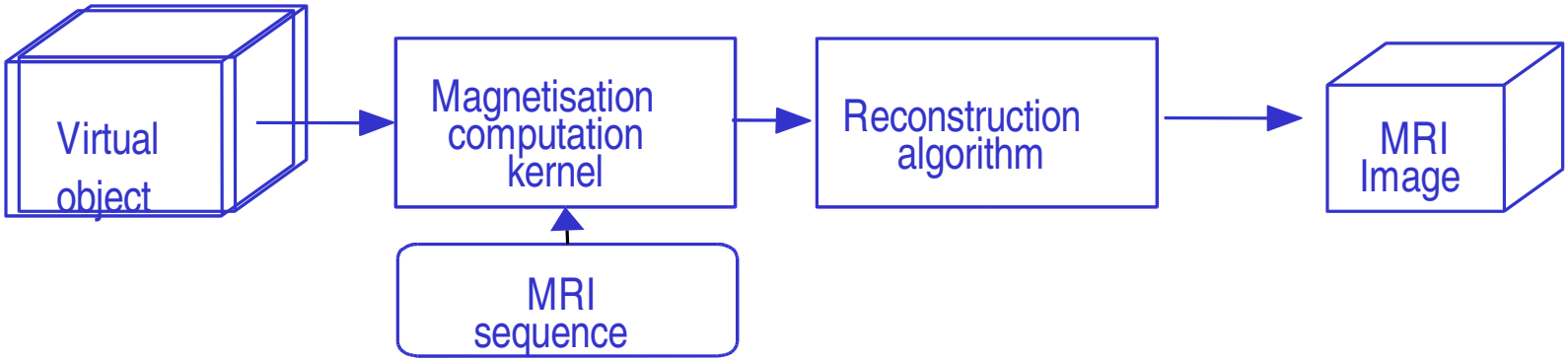
Most similar images



Low score images

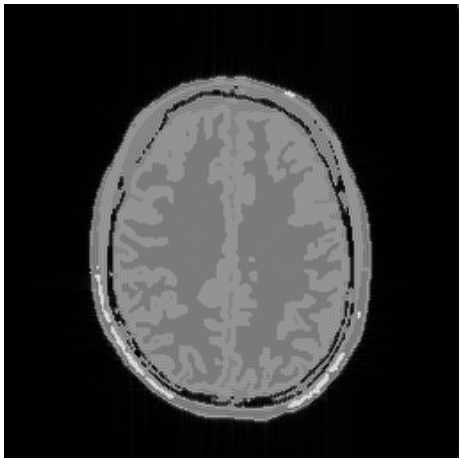
MRI simulation

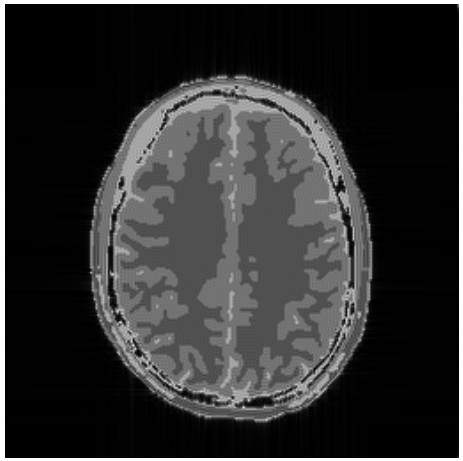
- Medical Resonance Image physics simulation



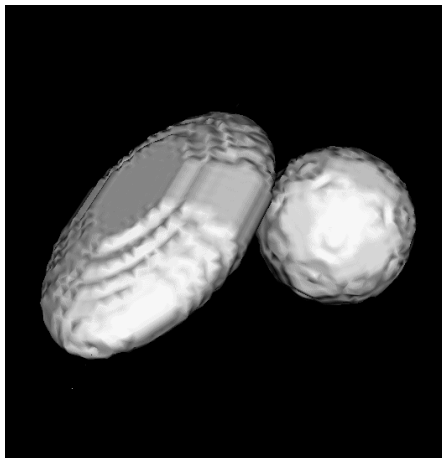
- Examples

2D (256^2) brain MRI





3D (64^3)



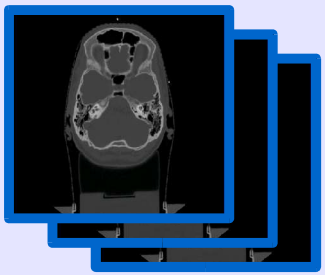
MRI simulation

- Kernel parallelization using MPICH-G2
- Processing time (8 processors)

2D: Image size	32^2	64^2	128^2	256^2	512^2	1024^2
Time	0.9s	3.4s	43.1s	12mn	201mn	3277mn
3D: Image size	16^3	32^3	64^3	128^3		
Time	4.9s	3.5mn	210mn	1626mn		

- Technical problems for large scale simulations

Monte carlo simulation for radiotherapy planning



Scanner slices: DICOM format

Concatenation

Image: text file

Binary file: Image.raw Size 19M

Anonymisation

Database

Retrieving of root output files from CEs

the CE

User interface

GATE

Computing Element

Storage Element

CCIN2P3

RAL

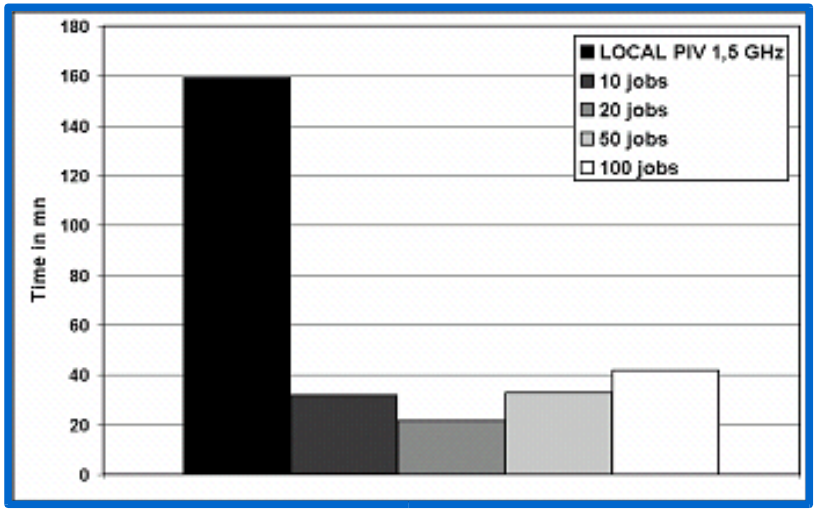
NIKHEF

MARSEILLE

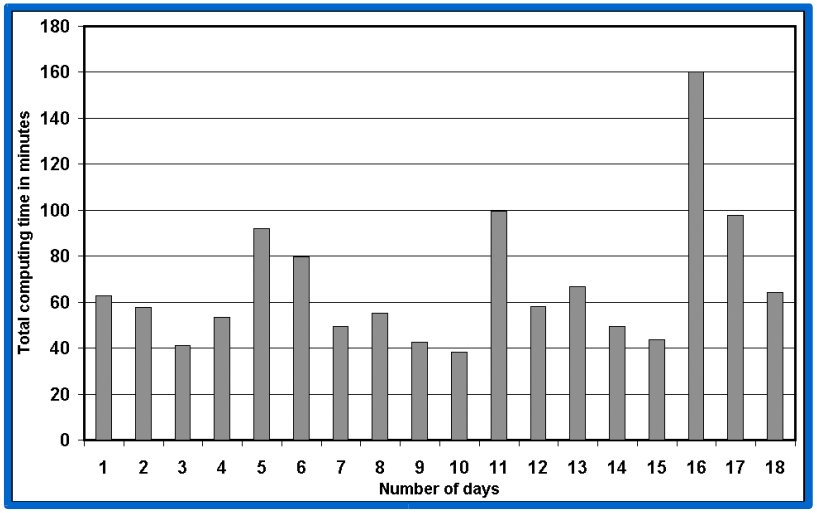
Monte carlo simulation for radiotherapy planning

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

Ocular brachytherapy simulation with 10M of events



Comparison of computing time



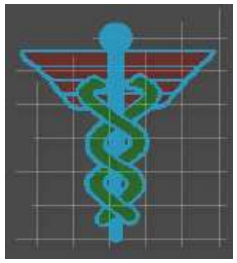
Parallel submission of 100 jobs

- It is not sufficient for clinical routine
- Necessary improvements
  - Dedicated resources (job prioritization)
  - Graphical User interface

Enabling Grids for E-science in Europe

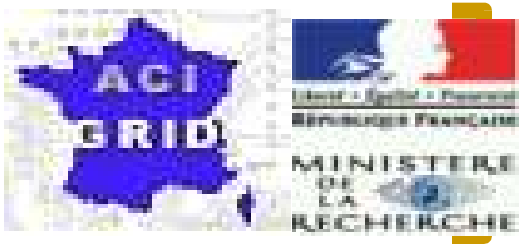
EGEE: Enabling Grids for E-science and industry in Europe

- Production platform
  - Production testbed deployment (EDG: up to 1000 CPUs, 10 sites, 15 TB disk space)
  - Real scale applications
- NA4 package: Applications
  - High energy physics
  - Biomedical applications
  - Other applications



MediGrid

http://www.creatis.insa-lyon.fr/MEDIGRID



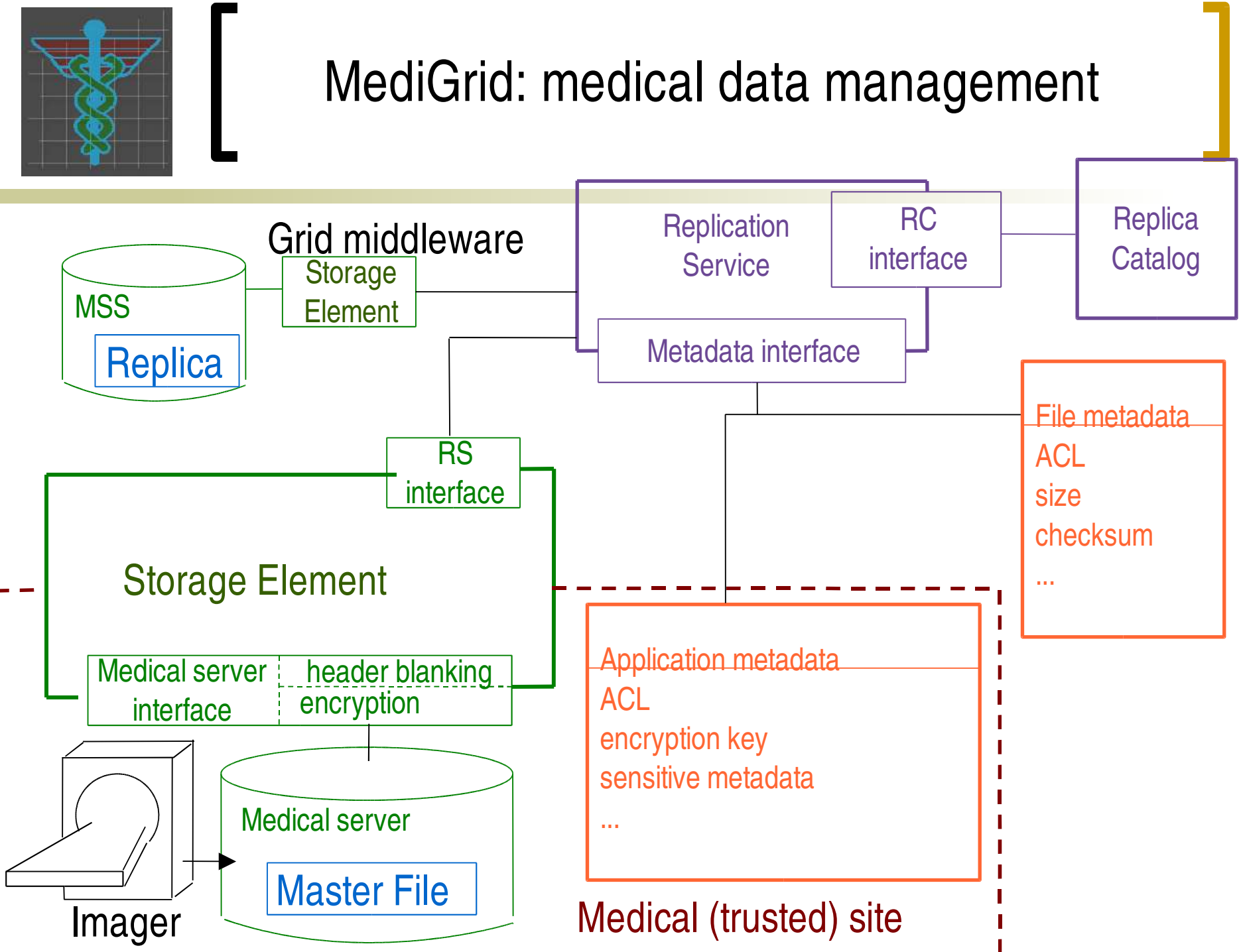
- Two laboratories in life science (Inserm) and computer science (CNRS)

Creatis LIRIS CNRS Inserm

Institut national
de la santé et de la recherche médicale

- Use computation GRIDs to face recent challenges in medical data analysis. We are focusing on two application kinds:
  - Computation intensive image processing algorithms
    - Parallelization
    - Reduced computation time
  - Management of very large datasets
    - Distributed storage
    - Massive distributed processing
    - Statistical analysis

MediGrid: medical data management



Conclusions

- Proof of concept level
  - Small scale demonstrator
- Need for large scale applications
  - Community awarness, real impact demonstrator
- Scientific and technical issues remain
  - Security...
  - Heterogeneous data format...
- Applications to be adapted for a new architecture
  - Change existing applications design
  - Identify new applications
- Acceptance by the medical domain
  - Interface
  - Reliability