Grid computing for medical image processing
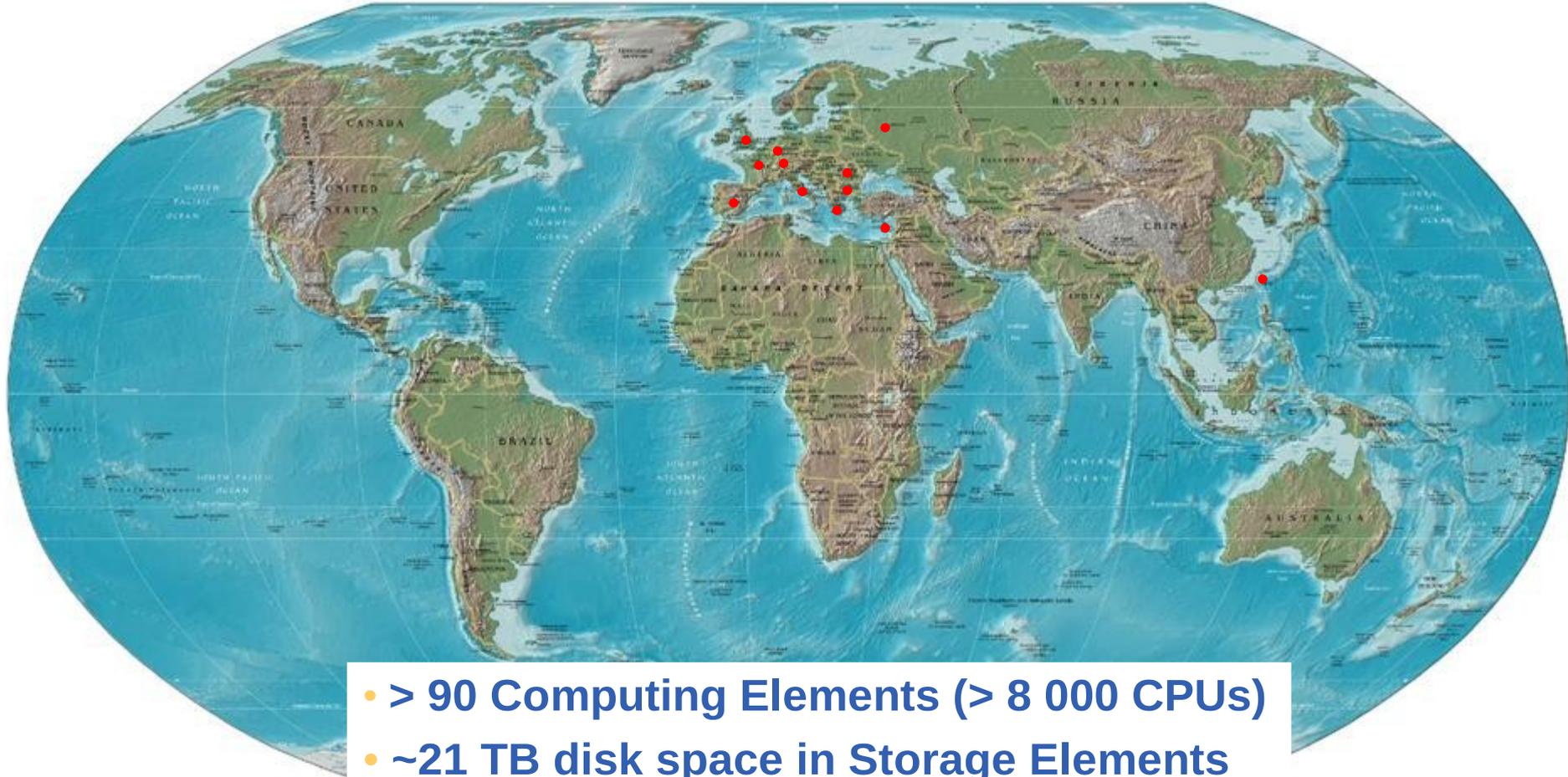
Johan Montagnat

GARR conference "Health Science Community"

Roma, December 12, 2007

www.eu-egee.org

Information Society
and Media

- **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
- **EGEE-I (2004-2006): NA4/biomed: pilot application area**
  - 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

- **> 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
  - ...

Enabling Grids for E-sciencE

- **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**
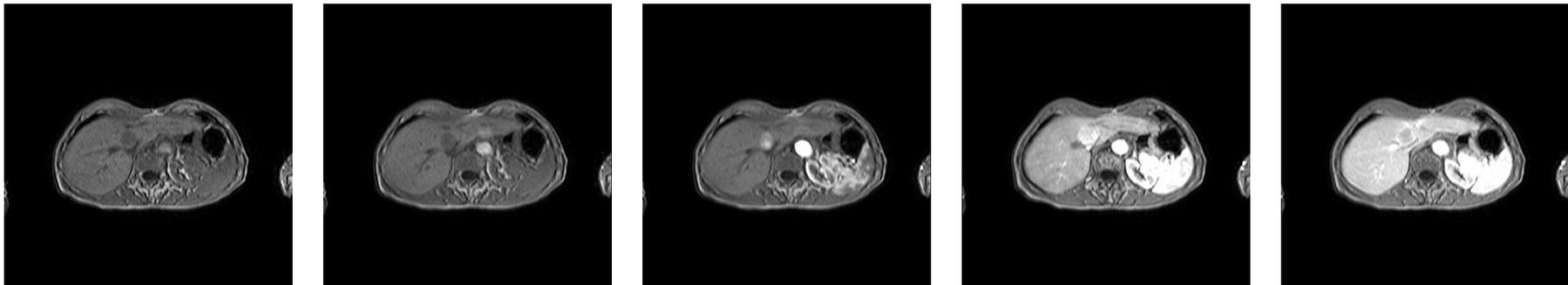 - I3S (CNRS – Sophia), INRIA

- **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.
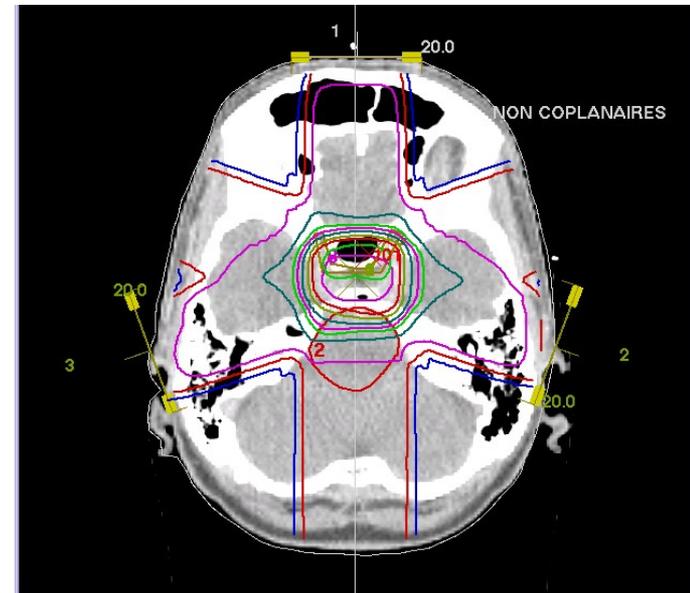
- **Scientific objectives**

  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.
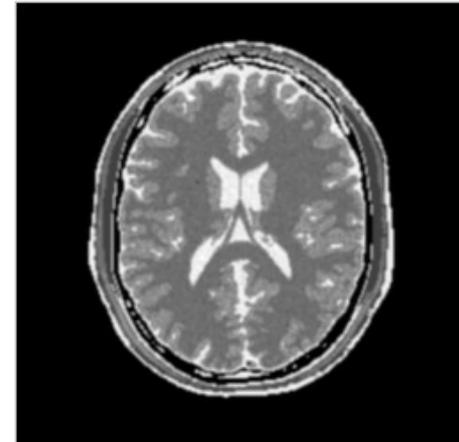
- **Method**

  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)

- **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.

- **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.
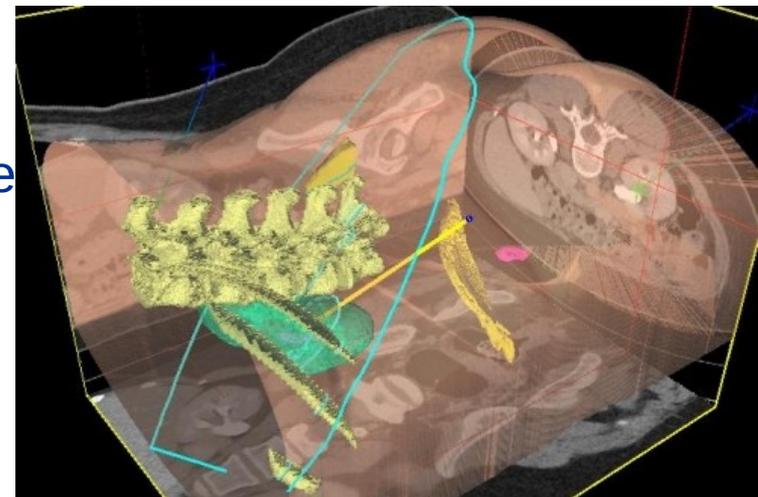
- **Method**

    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
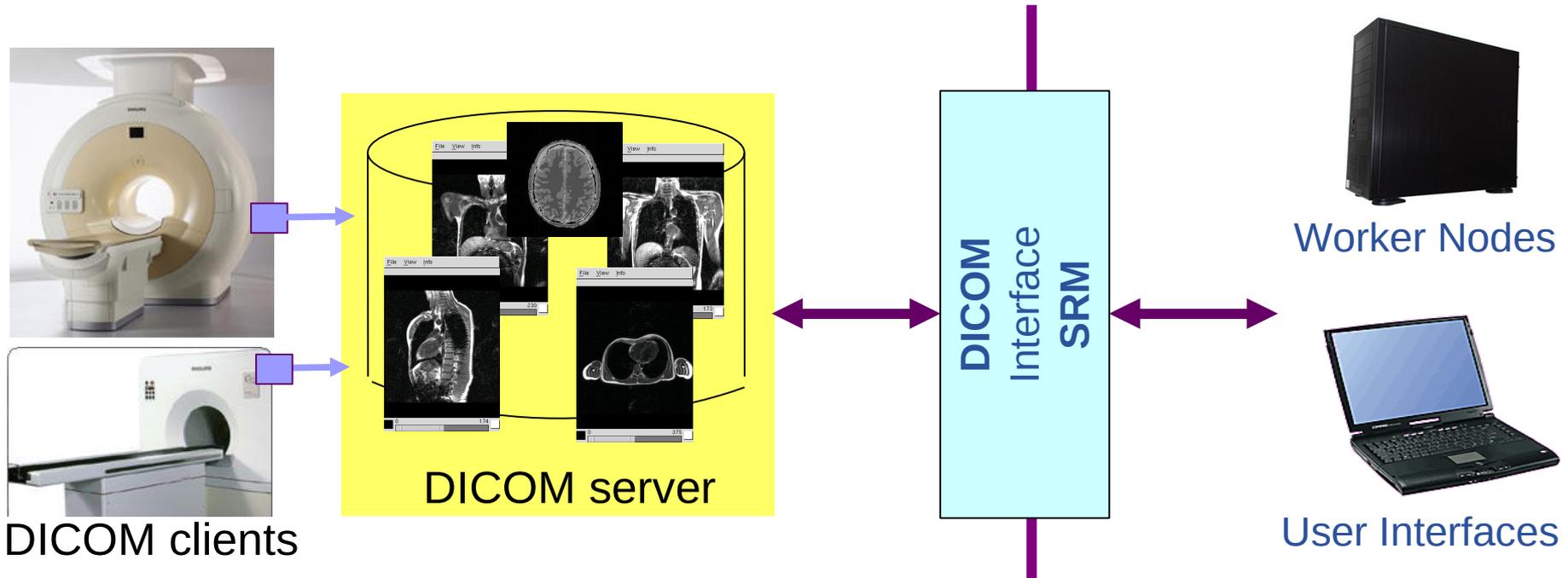   slices.

- **"Embarassingly" parallel problem**
- **Fast turn over of short jobs**
- **Data confidentiality**
- **Interactivity**
- **Fine grain parallelism (MPI)**
- **Workflow-based**
- **Portal**

Enabling Grids for E-sciencE

- **Medical Data Manager**
  - Grid storage interface to DICOM databases
- **Data protection**
  - Encryption (storage/network protection) + access co
    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

Medical Data Manager

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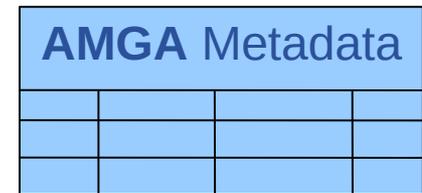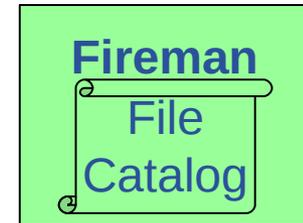 - Do not interfere with clinical practice

DICOM clients

DICOM server

DICOM Interface SRM

Worker Nodes

User Interfaces

- **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 anonimization**
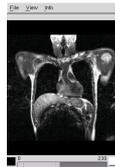   - It is based on the dCache implementation (SRM v1.1)
- **Data protection**
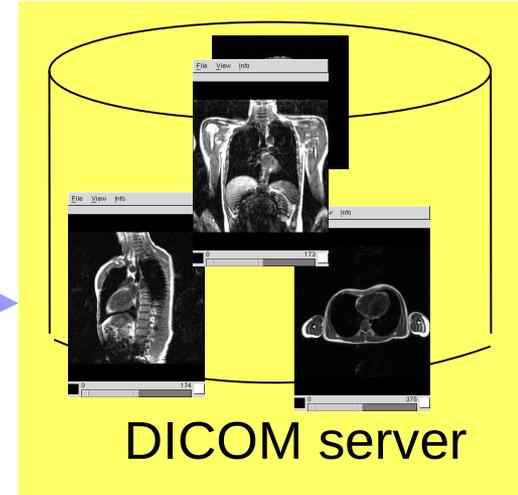 - **Hydra** provides encryption/ decryption **transparently**

Fireman
File Catalog

gLiteIO
server

AMGA Metadata

SRM-DICOM
Interface

Hydra
Key store

1. Image is acquired

2. Image is stored in DICOM server

3. glite-eds-put

DICOM server

gLiteIO server

3a. Image is registered

4. image metadata are registered

3b. Image key is produced and registered

Fireman File Catalog

AMGA Metadata

Hydra Key store

![egee logo] Enabling Grids for E-sciencE

1. get GUID from metadata → AMGA Metadata

Metadata DN-based ACL

User Interface

Worker Node

gLiteIO client

In-memory decryption

2. glite-eds-get

Fireman File

File DN-based ACL

gLiteIO server

3. get SURL from GUID

4. request file

7. get file key and decrypt file locally

6. on-the-fly encryption and anonimyzation return encrypted file

SRM-DICOM interface

Key DN-based ACL

Hydra Key store

5. get file key

Anonymization & encryption

~100 image pairs

~800 EGEE jobs

Enabling Grids for E-sciencE

- **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**
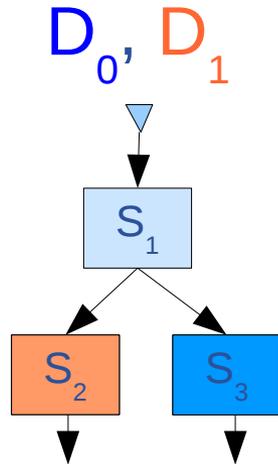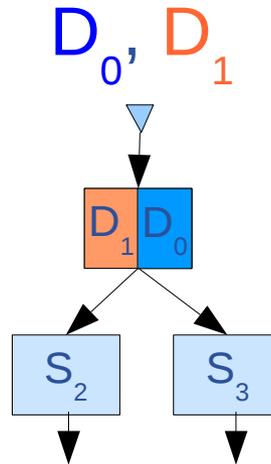 - Code and data parallelism

Enabling Grids for E-sciencE

- **A workflow naturally provides application parallelization**
- **MOTEUR transparently exploits 3 kinds of parallelism**

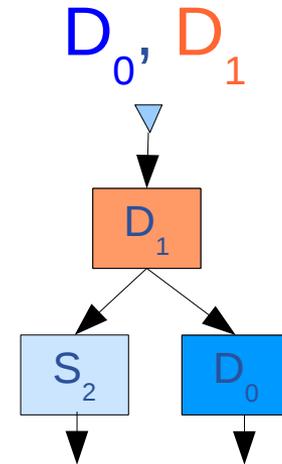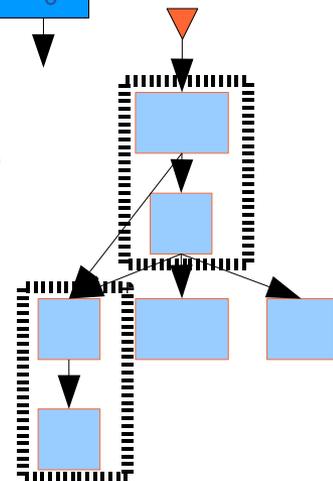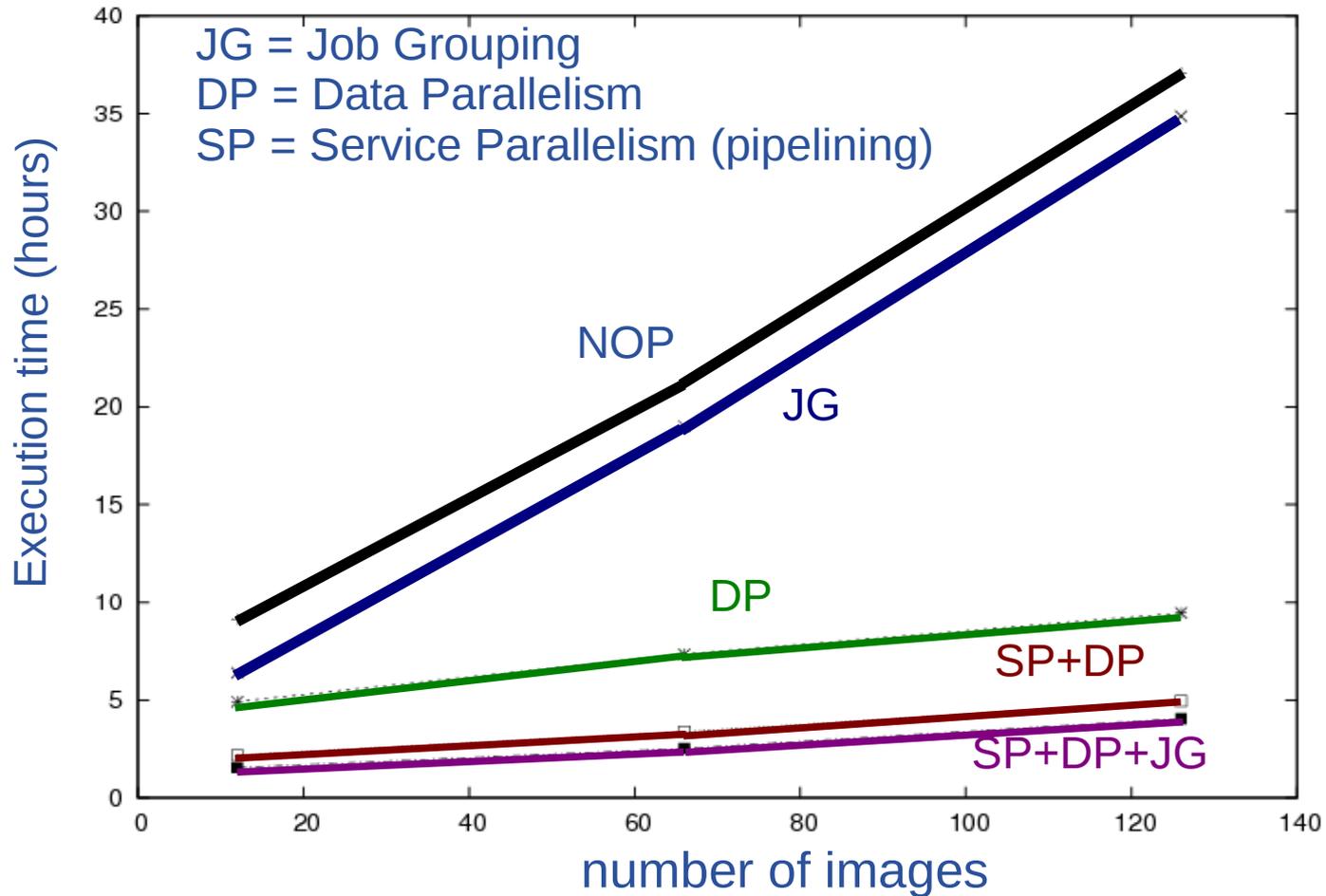*Workflow parallelism*   *Data parallelism*   *Service parallelism*

D_0, D_1 D_0, D_1 D_0, D_1

- **Jobs grouping strategy in sequential branches in order to reduce grid latency**

- **On the EGEE infrastructure (biomed VO)**

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

Enabling Grids for E-sciencE

- **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
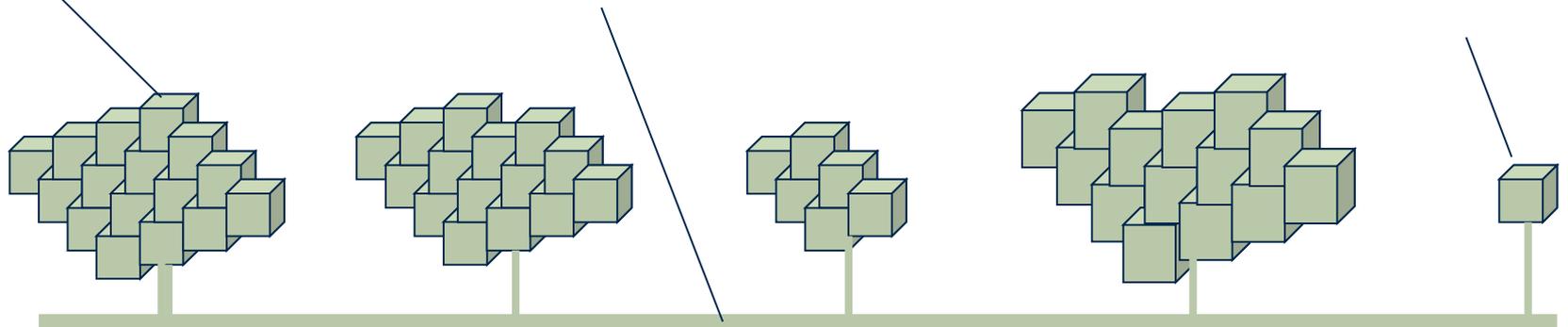 - Access to queue also limited

Virtualization technologies for grid applications isolation and deployment
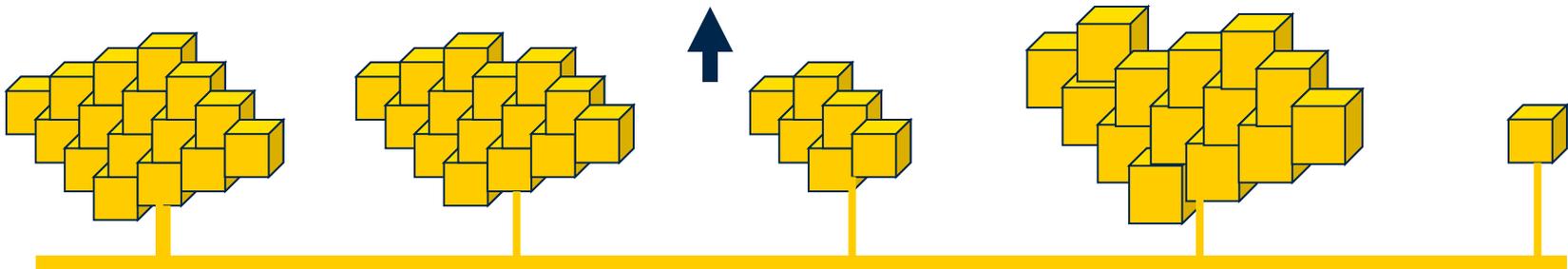http://hipcal.lri.fr

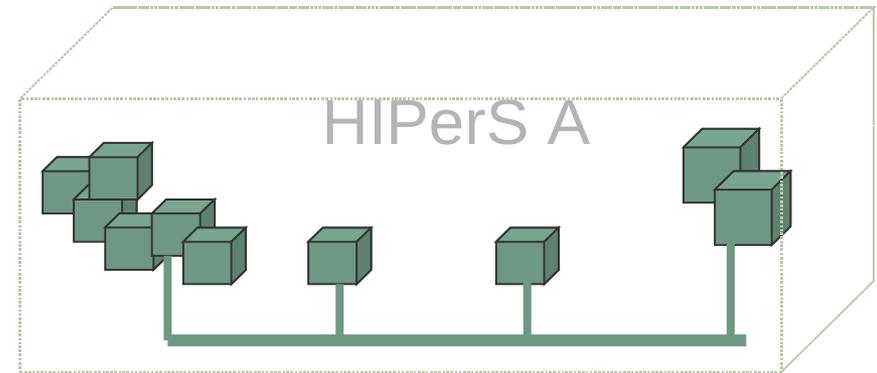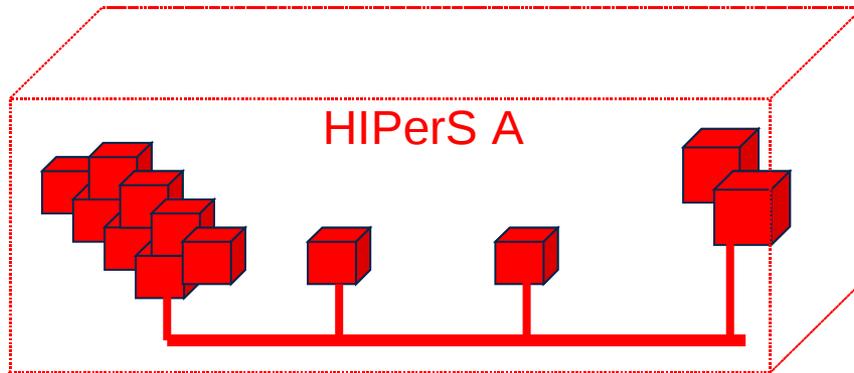cluster   Long distance network   Individual node

Given a Global infrastructure composed of
the aggregation of (computing) resources interconnected
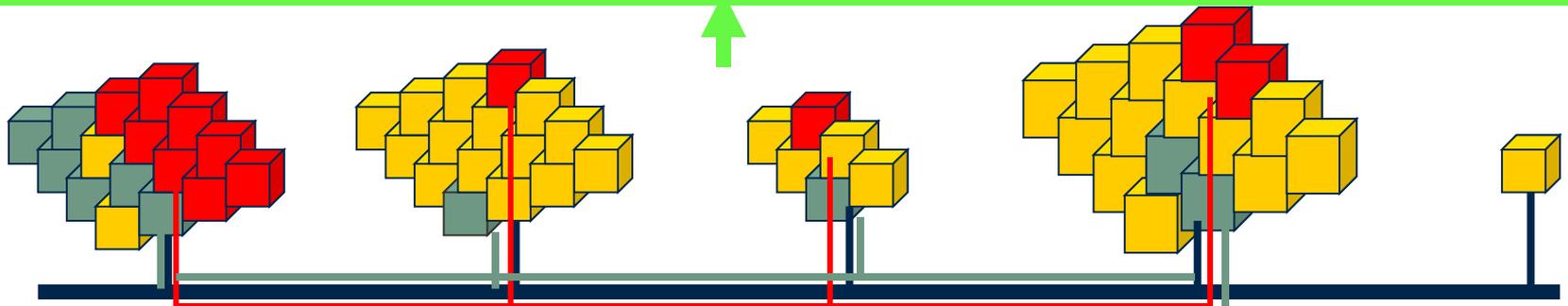by Internet or any network (ie: G5000 VPN)

HIPerSpace

Transform the Global infrastructure into an HIPerSpace:
collection of HIPerNodes and HIPerLinks (virtualized resources)
registered in an HIPerNET registrar (P2P)

Enabling Grids for E-sciencE

HIPerS A

HIPerS A

HIPerNET engine: select, allocate, schedule nodes&channels

Global infrastructure

Enabling Grids for E-sciencE

- **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...)**