



Enabling Grids for E-science

Forward Planning & Open Discussion

C. Loomis (LAL-Orsay)

*EGEE'06 Conference (Geneva)
25-29 September 2006*

www.eu-egee.org



- **Talk contains my own questions and views regarding the UIG. Not in anyway “official”!**
- **Points are intended to spur discussion and may seem (be) provocative.**
- **Probably easiest to go through talk quickly and go through again for the discussion.**
- **Don't be afraid to voice your opinion.**

- **Short-term**
 - Definition of technical details for use cases.
 - Completion of first use cases.

- **Medium-term**
 - Extension of use cases.
 - Maintenance (verification) of existing use cases.
 - Identification of other types of important documentation.
 - How to make wider documentation coherently available?

- **Long-term**
 - Scope of the UIG documentation reviews?
 - How to make documentation infrastructure self-sustaining?

- **Use cases:**

- Are the defined use cases sufficient for now?
- How to gather feedback for evolving the use cases?
- Necessary to keep all versions of use cases?
- Notification of new/updated use cases?
- Mechanism for unifying use cases?

Beginner	Typical	Skilled
Get certificate	Resource/service discovery	Software installation
Run a job	Jobs with data requirements	Large-scale data transfer
Copy/register/access files	Environment setup	Monitoring status (R-GMA)
Recovering results	Monitoring status	Data encryption
Monitoring job status	Software installation	AMGA metadata
Preparing a job	Short-deadline job submission	MPI
		Workflow examples
		VO deployed services
		Biomed app. kernel
		Geo. app. kernel

- **Use Cases:**

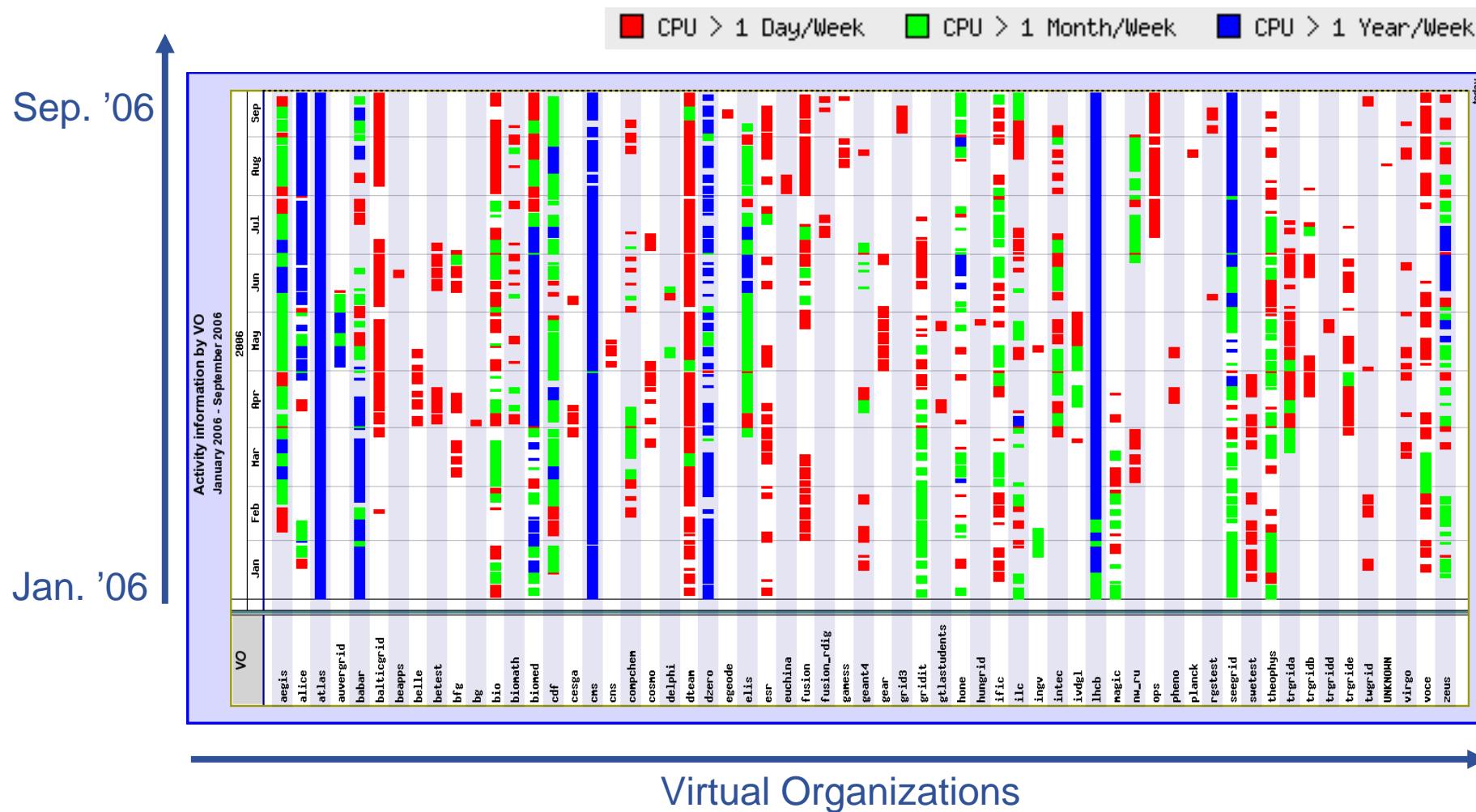
- Who's involved in defining new use cases? Free-for-all?
- How to manage the updates of use cases?
- How to systematically validate use cases with releases?

- **General Documentation:**

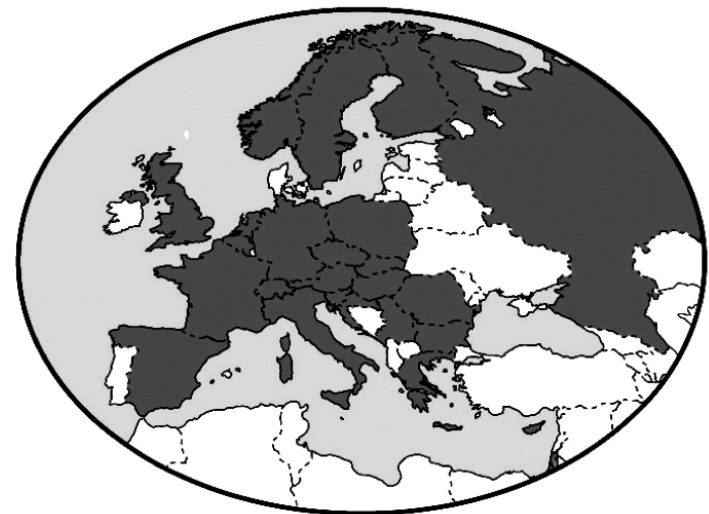
- What documentation is needed? How to determine this?
 - § gLite
 - § non-gLite services
 - § administrative procedures
- How to index/organize all of the user documentation?
 - § web site
 - § document repository
 - § search engine

- **Scope of UIG activities:**
 - Should UIG limit itself to strictly gLite documentation?
 - Expand to commonly used third-party services?
- **Review of documentation:**
 - “Bug” tracking for tracking documentation deficiencies
 - Informal review of documentation
 - Formal review of documentation
- **Sustaining the documentation infrastructure?**
 - Is this important?
 - How to make it self-sustaining?

- **Routine and large-scale use of EGEE infrastructure to produce scientific results.**
- **VOs:**
 - 165+ VOs (90+ registered) using the grid
 - App. Deploy. Plan (<https://edms.cern.ch/document/722131/2>)
- **Domains:**
 - **High-Energy Physics:** LHC, Tevatron, HERA, ...
 - **Biology:** Medical Images, Bioinformatics, Drug Discovery
 - **Earth Science:** Hydrology, Pollution, Climate, Geophysics, ...
 - **Astrophysics:** Planck, MAGIC
 - **Fusion**
 - **Computational Chemistry**
 - **Related Projects:** Finance, Digital Libraries, ...
 - **New areas:** nanotechnology, ...



- **Application Identification and Support (NA4)**
 - 25 countries, 40 partners, 280+ participants, 1000s of users
- **Support the large and diverse EGEE user community:**
 - **Promote dialog:** Users' Forums & EGEE Conferences
 - **Technical Aid:** Porting code, procedural issues
 - **Liaison:** Software and operational requirements
- **Need active participation:**
 - **Feedback:** Infrastructure, configuration, and middleware
 - **Resources:** Hardware and human



- **Evolution of Project (2001–now):**
 - European DataGrid: R&D
 - EGEE: Re-engineering & Infrastructure
 - EGEE-II: Infrastructure & Re-engineering
- **Evolution of Grid Users:**
 - **Focus:** Grid technology \Rightarrow Scientific results
 - **Goal:** Grid technology \Rightarrow Grid as a tool
 - **Experience:** IT experts \Rightarrow IT “minimalists”
- **These changes are healthy, but...**
 - Rely less on IT competence of users.
 - More portable, more flexible middleware.



larger grid
↓
more apps.



- **Simulation**
- **Bulk Processing**
- **Responsive Apps.**
- **Workflow**
- **Parallel Jobs**
- **Legacy Applications**

- **Examples**

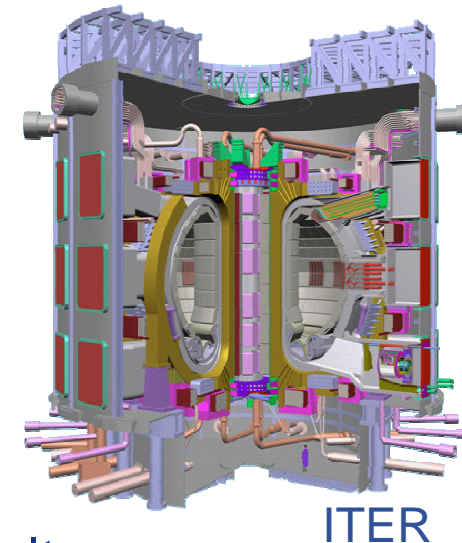
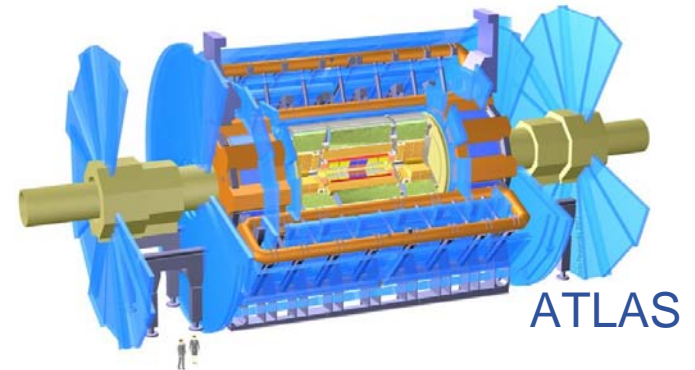
- LHC Monte Carlo simulation
- Fusion
- WISDOM—malaria/avian flu

- **Characteristics**

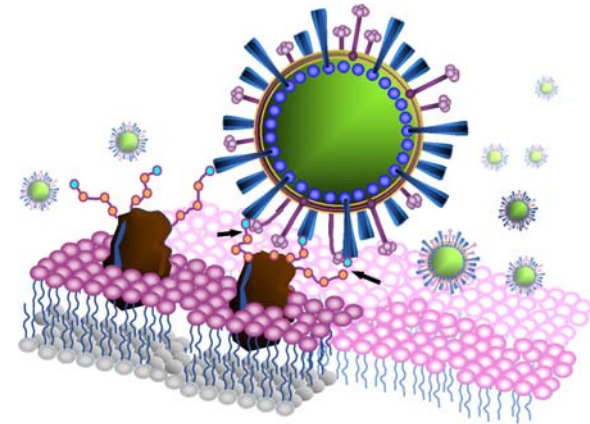
- Jobs are CPU-intensive
- Large number of independent jobs
- Run by few (expert) users
- Small input; large output

- **Needs**

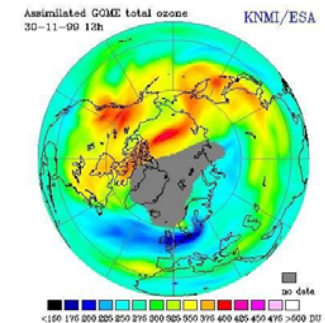
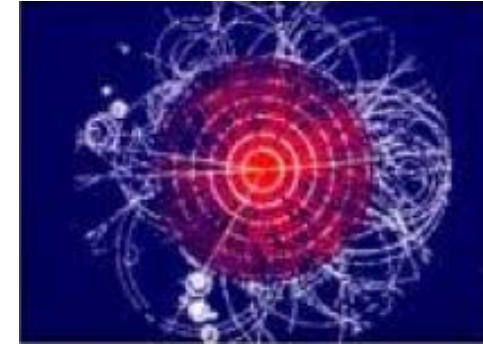
- Batch-system services
- Minimal data management for storage of results



- **WISDOM focuses on in silico drug discovery for neglected and emerging diseases.**
- **Malaria — Summer 2005**
 - 46 million ligands docked
 - 1 million selected
 - 1TB data produced; 80 CPU-years used in 6 weeks
- **Avian Flu — Spring 2006**
 - H5N1 neuraminidase
 - Impact of selected point mutations on eff. of existing drugs
 - Identification of new potential drugs acting on mutated N1
- **Fall 2006**
 - Extension to other neglected diseases



- **Examples**
 - HEP processing of raw data, analysis
 - Earth observation data processing
- **Characteristics**
 - Widely-distributed input data
 - Significant amount of input and output data
- **Needs**
 - Job management tools (workload management)
 - Meta-data services
 - More sophisticated data management



- **Examples**

- Prototyping new applications
- Monitoring grid operations
- Direct interactivity

- **Characteristics**

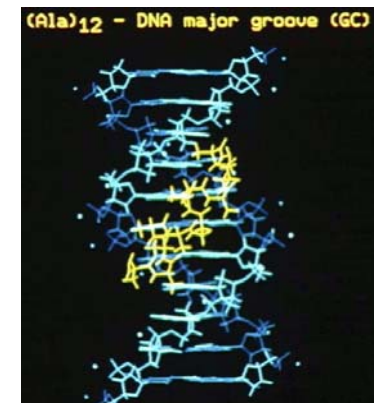
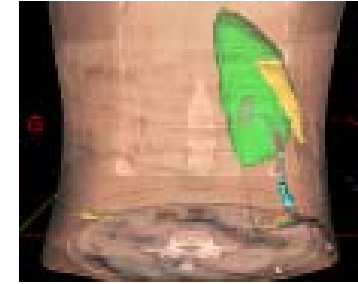
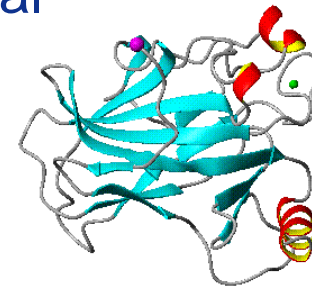
- Small amounts of input and output data
- Not CPU-intensive
- Short response time (few minutes)

- **Needs**

- Configuration which allows “immediate” execution (QoS)
- Services must treat jobs with minimum latency

- **Grid as a backend infrastructure:**

- gPTM3D: interactive analysis of medical images
- GPS@: bioinformatics via web portal
- GATE: radiotherapy planning
- DILIGENT: digital libraries
- Volcano sonification



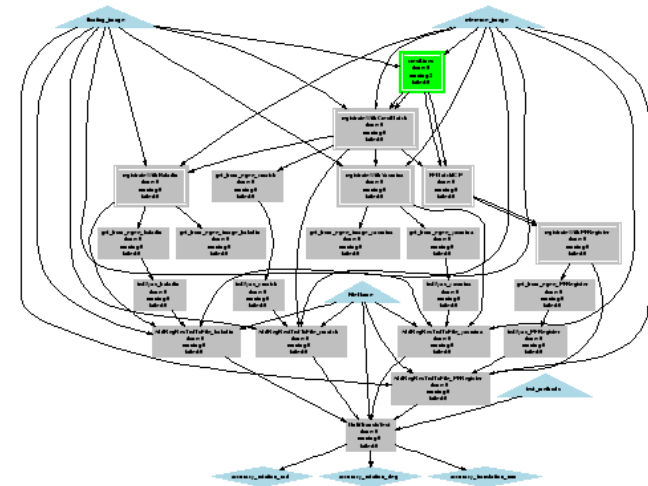
- **Characteristics**

- Rapid response: a human waiting for the result!
- Many small but CPU-intensive tasks
- User is not aware of “grid”!

- **Needs**

- Interfacing (data & computing) with non-grid application or portal
- User and rights management between front-end and grid

- **Examples**
 - “Bronze Standard”: image registration
 - Flood prediction
- **Characteristics**
 - Use of grid and non-grid services
 - Complex set of algorithms for the analysis
 - Complex dependencies between individual tasks
- **Needs**
 - Tools for managing the workflow itself
 - Standard interfaces for services (I.e. web-services)



- **Examples**

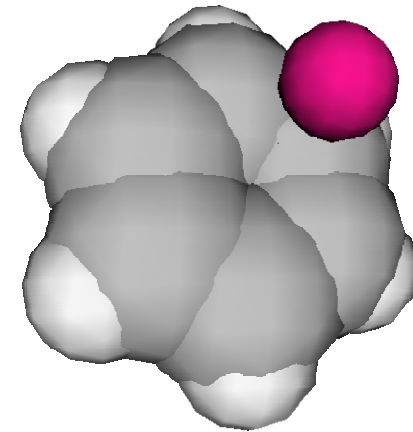
- Climate modeling
- Earthquake analysis
- Computational chemistry

- **Characteristics**

- Many interdependent, communicating tasks
- Many CPUs needed simultaneously
- Use of MPI libraries

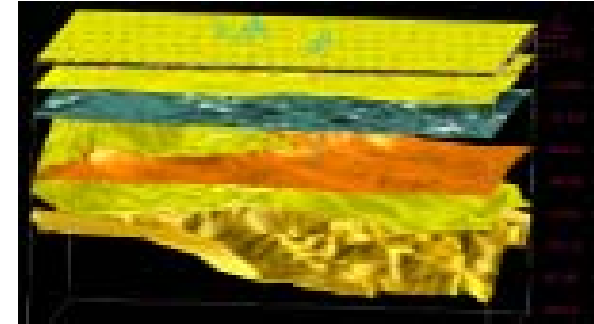
- **Needs**

- Configuration of resources for flexible use of MPI
- Pre-installation of optimized MPI libraries



- **Examples**

- Commercial or closed source binaries
- Geocluster: geophysical analysis software
- FlexX: molecular docking software
- Matlab, Mathematics, ...



- **Characteristics**

- Licenses: control access to software on the grid
- No recompilation \Rightarrow no direct use of grid APIs!

- **Needs**

- License server and grid deployment model
- Transparent access to data on the grid

- **Security**

- Ability to control access to services and to data
 - § Fine-grained access control lists
 - § Encryption & logging for more demanding disciplines
 - § Access control consistently implemented over all services

- **VO Management**

- Management of users, groups, and roles
- Changing the priority of jobs for different users, groups, roles
- Quota management for users, groups, roles
- Definition and access to special resources
 - § Application-level services
 - § Responsive queues (guaranteed, low-latency execution)

- **Services exist for many of the application needs and plans exist to fix existing deficiencies or holes.**
- **No longer “one-size-fits-all” world:**
 - Works for low-level services (CPU, storage).
 - Higher-level services imply trade-offs:
 - § E.g. latency vs. bulk response of meta-schedulers
 - § E.g. security vs. speed for data access
 - Commonalities allow “one-size-fits-many” solutions.
- **Future evolution:**
 - Standards more important than ever: plug-and-play services.
 - Diversification of higher-level services is healthy and inevitable.
 - Integration of third-party tools an absolute necessity.

- **Observe routine and large-scale use of the EGEE infrastructure by numerous, diverse set of users.**
- **EGEE provides backbone services which support wide range of different grid application families.**
 - Simulation, Bulk Processing, Responsive Apps., Workflow, Parallel Jobs, Legacy Applications
- **Third-party tools are becoming increasingly important for providing specialized (but flexible) services to particular groups of applications.**

- **Related projects:**
 - DEGREE
 - DILIGENT
 - EGRID
 - EU ChinaGRID
 - EU MedGRID
 - GRIDCC
 - many more...
- **Other collaborations:**
 - Geant4
 - ITU
 - ProActive
 - many more...

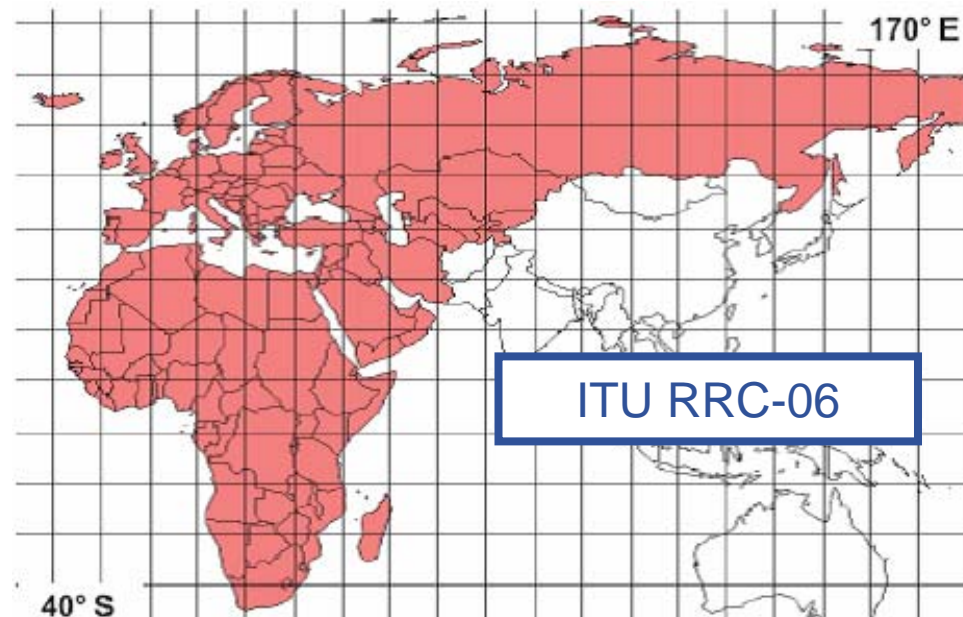


Figure 1
The extent of the planning area for the RRC-06

- **EGEE Conferences and Users' Forums**
 - Share your expertise, learn from other users.
 - Be open to collaboration with others.
- **Do (or don't) like something, speak up!**
 - VO issues, needs \Rightarrow VO Managers' Group
 - Resource, proc. problems \Rightarrow Operations Advisory Group (OAG)
 - Talk with NA4 steering committee
- **Report problems:**
 - Don't be afraid to use GGUS.
 - Report middleware annoyances \Rightarrow someone else is annoyed too!
- **NA4 website (<http://egeena4.lal.in2p3.fr/>)**