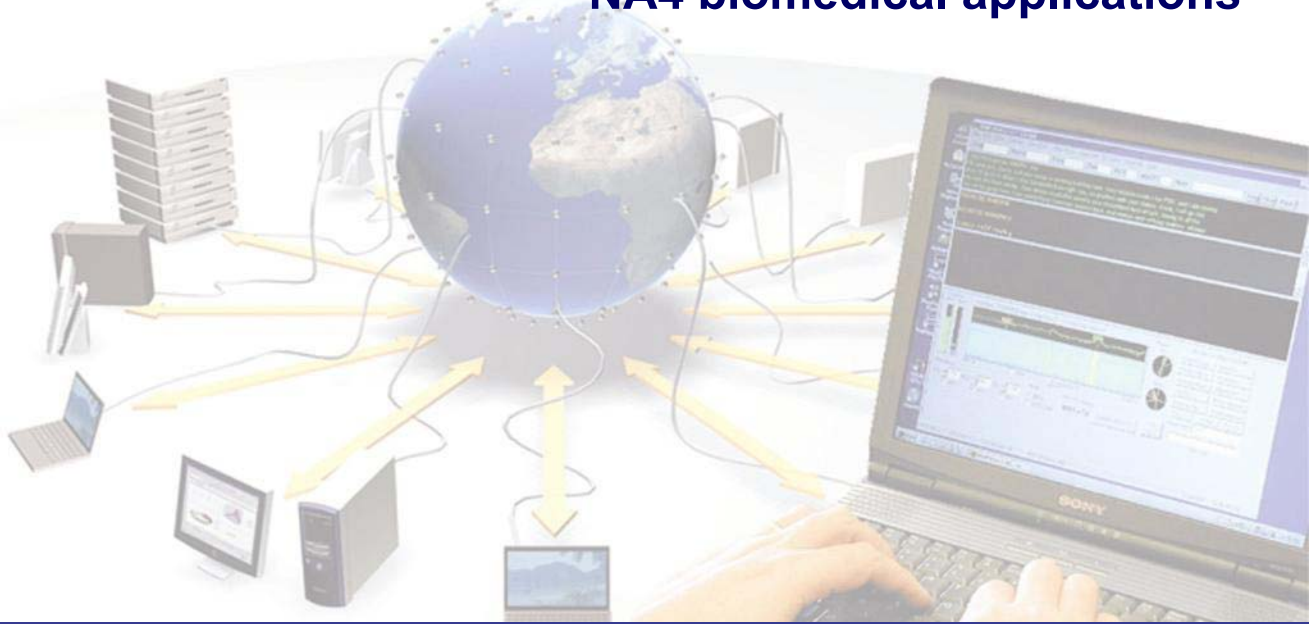
Enabling Grids for E-science in Europe

www.eu-egee.org

NA4-SA1 joint meeting, october 8th, 2004

GPSA Biomedical Demo

Vincent Lefort
NA4 biomedical applications



EGEE is a project funded by the European Union under contract IST-2003-508833

Contents

- GPSA overview
- LCG2 resources usage
- GPSA progress report
- GPSA demo roadmap
- Time to GPSA demo!

GPSA overview

- **GPSA : http://www.gpsa.fr**
- **To be an integrated grid portal devoted to molecular bioinformatics**
- Experience of porting the NPSA (Network Portein Sequence Analysis) services onto the EGEE grid.
- NPSA : http://npsa-pbil.ibcp.fr
  - Production web portal hosting proteins databases and algorithmes for sequences analysis.
  - Online since 1998.
  - Currently, strong restrictions in terms of databanks and algorithmes due to limited resources (one cluster of 14 CPUs)
  - Therefore the number of users connecting to the portal and the size of the data sets are restricted by the server but they will have to process.
- The same user community will be eager to (transparently) use the grid version of the same service once it has proven to be as stable and as efficient as the original service.

GPSA goals

Enabling Grids for E-science in Europe

- Deploying bioinformatic applications on the EGEE grid
- Made it available to the bioinformatic community
- Through an easy-to-use and efficient interface
  - Using the existing NPSA web portal interface
  - Well-known by the bioinformaticians and biologists
  - More than 5.000.000 since 1998
  - 3.000 analyses per day

LCG-2 resources usage (1)

- Sept 2004: GPSA (http://gpsa.ibcp.fr)
  - CPU: 12 CPUs (Local resources)
  - IO file size: from kB to GB
  - file storage: 10 GB used (Local resources)
  - NB of jobs: around 3000 per days, more than 5.000.000 since 1998 (NPSA's statistics)
  - Users: They are not logged, so it's difficult to know how many there are Certainly several thousands from NPSA statitics:
    - France 24 %
    - Europe 35 %
    - US 27 %
    - others 14 % Near future
- LCG-2 resources:
  - One User Interface « genomics-ui.ibcp.fr »
    devoted to the GPSA web portal (missing in user guide about CA certs)
  - Several files and DBs put on the Replica Manager Architecture
  - Launching jobs for demo preparation

LCG-2 resources usage (2)

Enabling Grids for E-science in Europe

- Future LCG2 resources usage
- feb 2005: EU review - GPSA (http://gpsa.ibcp.fr)
  - CPUs: 7-10 sites with 5-10 CPUs each and short queues
  - IO file size: from kB to GB
  - file storage: 500 Gb
- Far Future - late 2005: GPSA (http://gpsa.ibcp.fr)
  - CPUs: 10 sites with 20 CPUs each and short queues (see estimation remarks below)
  - IO file size: from kB to GB
  - file storage: 2-3 To (see estimation remarks below)
  - NB of jobs: around 3500 per days (see estimation remarks below)
  - Users: Users will be logged (see estimation remarks below)

LCG-2 resources usage (3)

- Estimation remarks
  - Currently, because of our lab limited resources, we have put several strong restrictions on the bioinformatic data and algorithmes available into the NPSA portal
  - So it's difficult to measure the potential user community of GPSA (Grid bionformatic portal).
  - We have to bring the application into a stable, efficient (short queues) and public state on the EGEE grid, suppressing major limitations (see before).
  - Then we will have good arguments to demonstrate the utility of the EGEE grid for the bioinformatic community.
  - And certainly the users will bring back others algorithmes and data to put on the grid.
  - This will refined in an iterative loop: it means that more than huge resources we need, during the next months, strong interactions with SA1 and resource centers to adapt the available resources with short reaction time.

GPSA progress report: LCG2 services usage

- Job submission by the portal: in progress
  - successfull validation of representative applications: **Done**
  - Short jobs (less than one minute) need short queues: **to be done (solution from Cal to be tested)**
  - Job collections (one JDL): **to be done**
- Data management: in progress
  - Using Replica Manager: **in progress**
  - Needing LFN<->local file substitution: **to be done (GFAL -> No)**
- Security (MWSG)
  - Authorizing « portal » certificate (submitting on behalf of non-registered users, *i.e*. without certificate): **to be done**
  - User authentication: **to be done**
  - Transfer and data encryption : **to be done**
- User interface:
  - Integration into the current NPSA portal: **to be done**
    - Transparently for the user
    - Made public and disseminate through the NPSA portal and community

GPSA Progress report: identifying « speeding down services »

- Logging GPSA’s jobs
- From submission on the portal to the visualisation of the results
- Identifying slow services that should be speeded-up in order to be relevant for bioinformatic application
  - Brokering
  - Job queuing (short queues)
  - Data access
    - registration with RM
    - smart brokering near SE
    - « open » access to SE)
  - …

GPSA demo roadmap

- October: use RLS to distribute replicate biomed databases on several SE. GFAL does not match our requirements for accessing registered data.
- Den Haag: test short queues CEs (performance). Demonstration scenario. Deploy on 5 sites with 5 to 10 CPUs each (CNB, UPV, CC-IN2P3...).
- December: Improve results visualization. Make some scale testing.
- Until review: Test compound jobs. Add processing algorithms.

Current GPSA demonstration

- Time to demo !
- And current demo online:
  - http://gpsa.ibcp.fr