Enabling Grids for E-sciencE

BioTutorial Introduction:

biomed scientists and the grid

NeSC Training Team

26-27th July 2005

www.eu-egee.org

Information Society

Outline

- **Purpose:**
  - To assist in the development of biomedical research

- **Achieved by:**
  - Matching the properties of grid to the properties of biomed problems

Properties of Grids

- **What can grids offer:**
  - Computational power
    - Access to computational resources which cannot be recruited in any other way
  - Data Federation
    - Bring many data resources together to be used as one.
  - Collaborative environments
    - New ways of collaborative working, including shared resources.

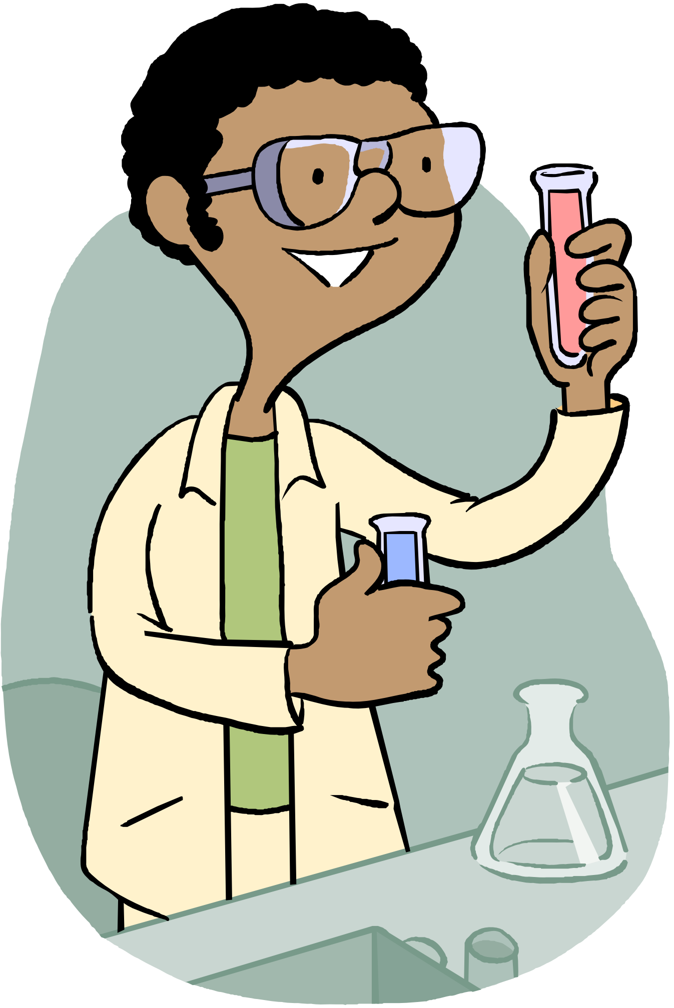
Distributed processing

- **A good problem for distributed processing:**
  - Can be readily split into many sub-tasks.
  - Ideally these sub tasks will be independent of each other.
  - Each of these tasks should require sufficient processing to balance the inherent network lag.
  - Ideally sub-tasks can be dealt with as batches.

- **A poor problem for distributed processing:**
  - Difficult to split up
  - Any sub-tasks are closely dependent on each other
  - Only very large or very small tasks can be identified
  - A great deal of interactivity required (time critical).

Frank – A biomed researcher with a problem

- **Frank is a biomed researcher**
- **He has a problem**
  - But we don't necessarily know what it is.
- **Frank has an existing application**
  - Taking this as given – we are not in the business of algorithm development at the moment.
- **Frank has some existing data**
  - At least the existing public databases.



What do we want to do for Frank?

- We need to give him the processing power that his application requires
- We need to give him easy access to the data he needs for his application
- Frank has a local cluster (say 20 machines) and a local CONDOR pool (maybe 100 machines)
- He has full access to public databases but doesn’t have the resource to keep them up to date by downloading them all to keep up with their update cycle.

An example workflow

- We have already decided that Frank's local resources are not sufficient
- Imagine Frank is working for an agrichem business/dept and they want to develop new targets for agriculturally important parasites.
- We could imagine that
  - he wants to compare sequences all the sequences in the genomes of (using Smith-Waterman)
    - 6 commerically important agricultural species,
    - adding human/rat/drosophila,
    - then adding know parasite genomes (nematodes?).
  - He wants to compare the results against structures in PDB
  - and finally compare these to a combinatorial chemistry library

Coming to EGEE

- **Frank has decided that he is faced with a couple of choices:**
  - Find a supercomputing center he can work with
  - Join a Virtual Organisation and collaborate on a grid.

How can Frank’s problem be brought to the Grid?

- **Frank already knows what his workflow is.**
- **He knows the applications and data he needs.**
  - Smith-Waterman sequence comparison algorithm
  - Annotation transfer application
  - Nucleic Acid – Protein translation method
  - Structural comparison algorithm
  - Chemical docking application
  - Various data extraction and format translation applications
  - Genome databases
  - Protein databases
  - Structural databases
  - Chemical structures library

Getting a sequence comparison algorithm onto the grid

- Take one command line program (BLAST, MPSRCH, etc).
- Check the install requirements
- Create tar or RPM of what you need
  - You can install/compile on a worker node
  - If it is a commonly used program it may already be there
- You may need to send a database with it and set up the links (if it's not there already)
- Or you may have to develop a way to point to a managed grid resource (file)
  - Requires more alteration.
- Write a script which sends your query sequences in batches to the RB and collects the results together

Getting more advanced

- Once you have a basic system running you might want to develop a more complex pipeline
- Use CONDOR DAGMAN to run a workflow on the worker node
  - Eg. Sequ comparion -> structural comparison -> molecular docking -> annotation
- Use myGrid TAVERNA to run a workflow over a variety of nodes.

How will biocomputing succeed on the grid

- **The only way that biological computing can be successful on grids is:**
  - For biologists to use their imaginations to ask biological questions which cannot be answered using today’s technologies!
  - For biologists to find new ways of working together and sharing resources using new technologies.
  - Computing scientists may provide the technologies but they are not equipped to ask the right questions.