Minimum intrusion Grid

'GeneRecon': A User Case from Bioinformatics

Early experiences

Introduction

- Genetic factors in inheritable diseases
- GeneRecon
  - analysis of population genetic data
    - locate areas of a genome which may be linked to "genetic diseases" (cancer, diabetes II, etc)
  - based on a Markov-Monte-Carlo method
    - search through a very large parameter space
- Experiment: genetic or env. cause
  - requires thousands of GeneRecon runs with different input

Resource Requirements

- GeneRecon
  - single CPU application
  - written mainly in C but using Scheme
  - very CPU-intensive
  - one run: 1-7 days on a 'standard' computer
  - limited memory and disk requirements
- Experiment
  - 30000 independent GeneRecon runs
  - full parallelism
  - all in all: at least 82 CPU-years

Finding Resources

- Only 12 node cluster locally available
- Grid
  - suitable resources
  - software requirements (Scheme)
- NorduGrid ARC and MiG
  - Atlas Data Challenge 2
  - available resources
- Mostly MiG

Grid Preparations

- Access
  - certificates, software?
- Partition work into jobs
- Minimize external dependencies
  - Scheme, C
- Describe jobs
  - Resource Specification Language
- Manage and monitor thousands of jobs
  - scripts

mRSL Example

::RUNTIMEENVIRONMENT::

GENERECON-2.1.3

::EXECUTE::

$GENERECON_HOME/generecon run-1304.scm

::JOBNAME::

GeneRecon-1304

::INPUTFILES::

run-1304.scm

::OUTPUTFILES::

locus.txt +JOBID+/locus.txt

cluster.txt +JOBID+/cluster.txt

connectedness.txt +JOBID+/connectedness.txt

likelihood.txt +JOBID+/likelihood.txt

::CPUTIME::

345600

Runtime Environment

- Description of job execution environment
- C library versions
  - three sites, three versions :-(
- GeneRecon RE
  - path to correct version
    - simply use GENERECON_HOME variable
  - static input files
    - only need to (implicitly) transfer one input file
  - jobs are only handed out to suitable resources

Conclusion

- Experiment is only feasible with Grid
- Nearly 1000 jobs finished
  - 4 CPU-years in 2 weeks
  - Still about 29000 of the 30000 jobs left
- MiG consistently delivers the resources!
- Analysis of results just started
- Future
  - greater projects now possible
  - SSS, check-pointing