Using the BiG Grid HPC Cloud Infrastructure

Han Rauwerda
Wim de Leeuw
Timo M. Breit

MAD/IBU
Swammerdam Institute of Life Sciences
University of Amsterdam

Transcriptomics Introduction

MAD/IBU (group Timo Breit): transcriptomics

- Analysis of gene transcription: wet lab + dry lab.

Transcriptomics experiments:

- 10 to >100 samples; 60 x 10^3 to >20 x 10^6 datapoints / sample;
- Many research questions/ many experimental designs
- Diverse platforms, many different wet-lab approaches
- De facto standard tools for many different tasks
  - e.g. array design, mixed effect ANOVA, module finding, construction of (Bayesian) networks, assembly of de novo transcriptomes, etc.
  - Analyses explorative.

Set up Problem Solving Environments (PSEs) for Transcriptomics

Transcriptomics Problem Solving Environments

Requirements for Transcriptomics PSEs

- Transcriptomics PSEs must be able to invoke HPC resources;
  - Most tasks embarrassingly parallel
    - Computational different per experiment, resources must be scalable.
    - Functional needs: experiment & researcher
- Flexible interfacing needed between local and HPC environment;
- Easy installation of transcriptomics specific software;

	Grid	Cloud
Flexible interfacing	no	yes
Root privileges	no	yes
Scalability	yes	yes

BiG Grid HPC Cloud Beta testing

1 year of testing in the BiG Grid HPC Cloud Beta testing:

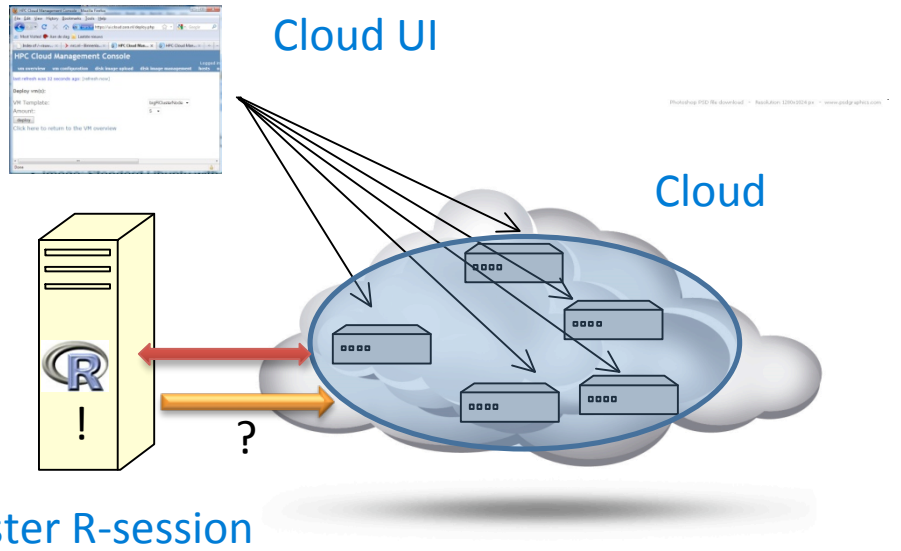
- Has become invaluable resource in daily research
- Used very frequently (over 8 times our quotum!)
- Stable middleware (Open Nebula)
- Support very good and well organized.

Worked on:

- interface to the cloud from a local environment
- cloud images to accommodate specific Problem Solving Environments
- usage of cloud images in education

Interfacing the cloud from a local environment

- Disk Image: standard Ubuntu with ssh & R
- Machine images with 4/8 cores, 2/4GB RAM
- External network
- Firewall exception, access using ssh-key
- How it works:
  - User has local R session
  - User starts VM's in Cloud UI (Nebula)
  - In R: poll cloud to recruit machines (1min.)
  - StartCluster()



The β testing - Results & Conclusion

1. Microarray analysis: *Calculation of F-values in a 36 * 135 k transcriptomics study using of 5000 permutations on 16 cores.*
   - worked out of the box (including the standard cluster logic)
   - no indication of large overhead
2. Ageing study - *conditional correlation*

   dr. Martijs Jonker (MAD/IBU), prof. van Steeg (RIVM), prof. dr. v.d. Horst en prof.dr. Hoeymakers (EMC)

   - 6 timepoints, 4 tissues, 3 replicates and 35 k measurements + pathological data
   - Question: find per-gene correlation with pathological data (staining)
   - Spearman Correlation conditional on chronological age (not normal)
   - p-values through 10k permutations (4000 core hours / tissue)

   Co-expression network analysis

   - 6k * 6k correlation matrix (conditional on chronological age)
   - calculation of this matrix parallellized. (5.000 core hours / tissue)
     - Development during testing period (real life!)
     - Many ideas were tried (clusters with 32 - 64 cores)
     - Cloud cluster: like a real cluster
     - Virtually no hick-ups of the system, no waiting times
     - User: it is a very convenient system

Cloud images to accommodate specific PSEs

- PSEs for array design
- PSEs for sequence alignment (non redundant GenBank)
- PSEs for module networks (Lemone)
- PSEs for Microsoft Windows applications (Genmapp)
- PSEs for Next Generation Sequencing Transcriptomics
  - de novo transcriptome assembly of mite with 80% identity
  - mapping of Illumina data on reference genomes
- Cloud in omics Education
  - prepare and adapt one image for a course
  - serve any number of students.

Conclusion

- Beta testing resulted in set of very useful PSEs
- Next gen sequencing: focus on cloud
  - (re-)use of EC-images
- Command line skills: biologists start to invest.
- Usage will soar: good accounting mechanism necessary
- Size of biological data increases:
  - bandwidth and storage buffers needed as access points to grid/cloud

Many thanks to the HPC cloud beta testing team!!

And looking forward to the new BiG Grid HPC Cloud