Galaxy on the GenomeCloud

Yet another on-demand Galaxy cloud, but only powered by Apache CloudStack

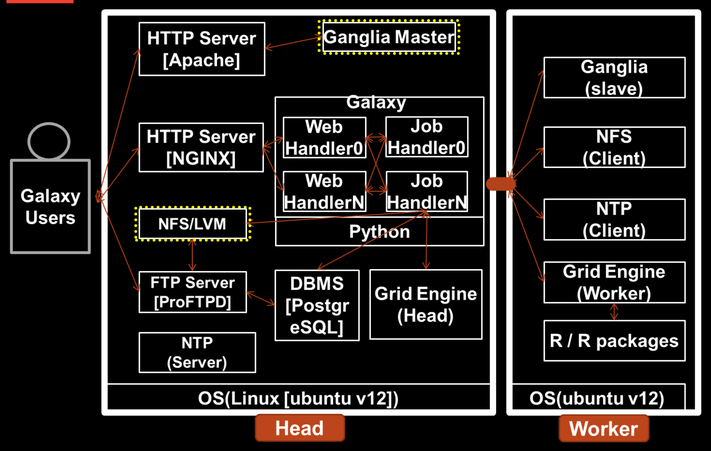
Youngki Kim, kt, Korea

Abstract

GenomeCloud is an integrated platform for analyzing, interpreting and storing genome data, based on KT's cloud computing infrastructure which uses Apache CloudStack software. GenomeCloud consists of g-Analysis (automated genome analysis pipelines at your fingertips), g-Cluster (easy-of-use and cost-effective genome research infrastructure) and g-Storage (a simple way to store and share genome-specific data).
Because of flexible tool integration architecture and seamless workflow creation functionality, Galaxy was selected to achieve multi purpose goals such as agile pipeline development and bioinformatics education support. To provide on-demand and Apache CloudStack based Galaxy cluster, we have automated virtual machine creation, clustering and various software setup including Galaxy.
Furthermore, seamless integration with GenomeCloud helps researchers not only create and manage Galaxy through a convenient web interface but also fully utilizes genome data in g-Storage. g-Storage is powered by OpenStack Swift and specially designed genome file transfer protocol. Galaxy on the GenomeCloud uses Grid Engine as a Cloud HPC Solutions, Ganglia as a distributed monitoring system and LVM over NFS as a large volume shared storage, all of which are setup automatically upon request.

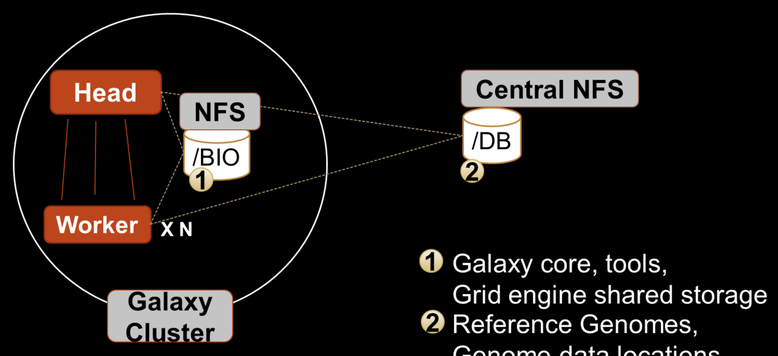
Galaxy on the GenomeCloud

Galaxy software stack



Galaxy system architecture

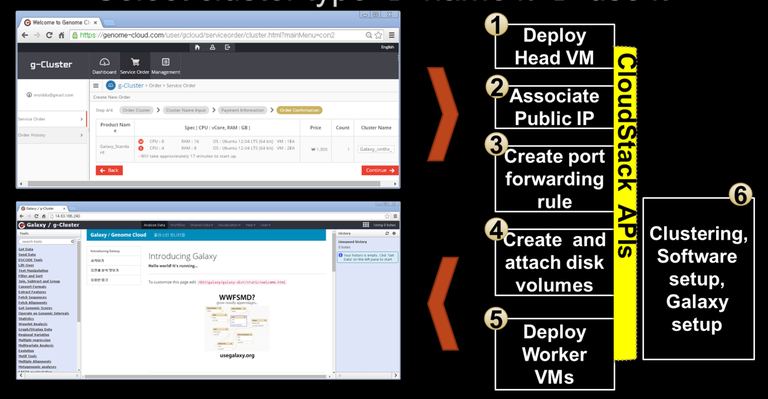
- Decoupled multi layer shared storages



1. Galaxy core, tools, Grid engine shared storage
2. Reference Genomes, Genome data locations

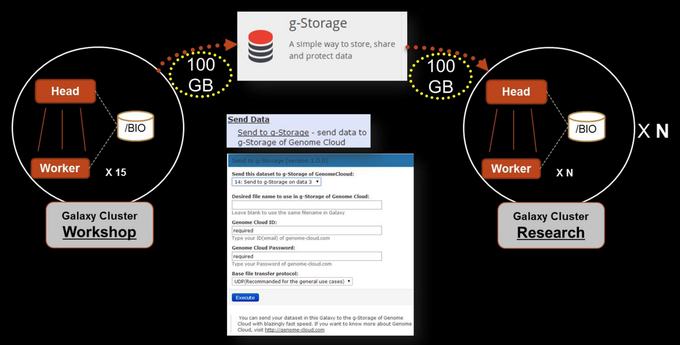
Fully automated cluster creation

- Select cluster type ➔ name it ➔ use it



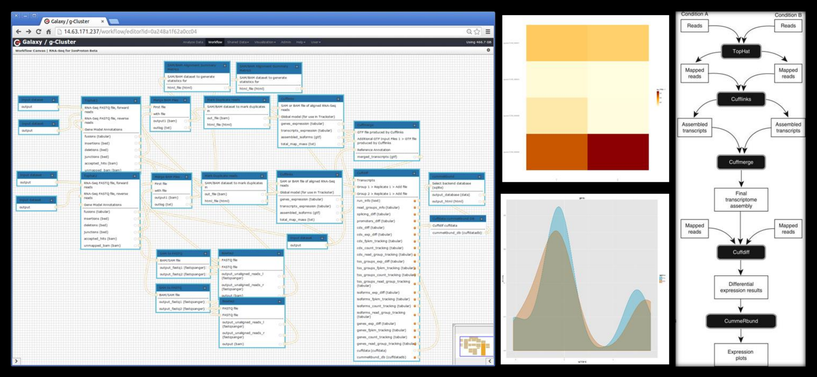
Integration with g-Storage

- Inter Galaxy cluster data transfer
  - Develop a galaxy tool for sending large data



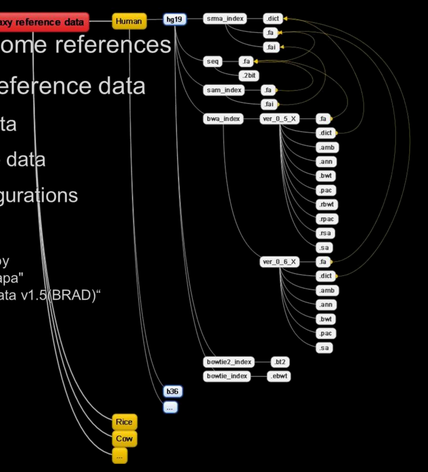
Pre-installed pipelines

- Workflows for RNA-Seq(Tuxedo, Ion Proton) analysis



Supporting diverse genome references

- Rapid support of new genome references
- A python tool for Galaxy reference data
  - Index reference genome data
  - Make directories and locate data
  - Change 10 + location configurations
  - Example:
    - python galaxy_reference_indexing.py -i "Brapa_sequence_v1.5.fa" -s "Brapa" -d "brapa_1.5" -e "Brapa genome data v1.5(BRAD)"

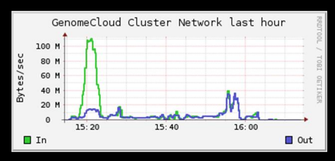


Use case & Acknowledgement

Bioinformatics education support

4 half-day workshops

- Dates: 13th, 20th May, 3rd,10th Jun
- **Attendees : 50 +**
- Galaxy : 8 core 16GB X 16 servers
- **Contents : from fastQC to RNA-Seq Analysis**
- Off site workshop and home works
  - # of executed jobs: 5,000 +
- Feedback: good to continue further research
- Lessons learned:
  - Be aware of bottlenecks
  - Fix it or adapt to it



Acknowledgment

- **Galaxy team**
  - ➔ We could never **start** this without you
- **GenomeCloud team**
  - ➔ We could never **finish** this without you
  - Daechul choi, Changbum hong, Kwangjoong kim, Wanpyo hong, Hankyu choi, Hosang jeon, Sehyuk yoon, Eunjean jo

GenomeCloud
http://genome-cloud.com

Galaxy Community Conference 2014, JHU, USA