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

Youngki Kim
kt, Korea
Outline

- **Introducing GenomeCloud**
  - Who we are, why we start
- Galaxy on the GenomeCloud
- Use cases and lessons learned
- Conclusions
Introducing GenomeCloud

Focus on your research, we do the rest

Watch the Video

http://genome-cloud.com

GenomeCloud

A complete and integrated platform from analyzing genome data to the interpretation of analysis results.

View all services

g-Analysis
Automated genome analysis pipelines at your fingertips.

g-Cluster
Easy-of-use and cost-effective genome research infrastructure

g-Storage
A simple way to store, share and protect data

g-Insight BETA
Accurate analysis and interpretation of biological meaning of genome data
Far away to the Galaxy

2X more hosts than GenomeCloud

- 5 X more time than GenomeCloud for data transfer
Outline

• Introducing GenomeCloud

• **Galaxy on the GenomeCloud**
  • Software stack, system architecture, automation and add-ons

• Use cases and lessons learned

• Conclusions
Launch GenomeCloud
2012.12

Start Galaxy setup in servers
• 2013.7

Complete Galaxy system architecture and software stacks
• 2013.9

Automated and integrated with GenomeCloud
• 2013.12

Launch Galaxy on the GenomeCloud
• 2014.01
Galaxy system architecture

- Decoupled shared storages

1. Galaxy core, tools, user data
   Shared storage for Grid Engine

2. Reference genomes data, Genome data locations
Fully automated cluster creation

- Select cluster type ➔ name it ➔ use it

1. Deploy head node
2. Associate public IP
3. Create port forwarding rule
4. Create and attach disk volumes
5. Deploy worker nodes
6. Clustering, Software setup, Galaxy setup
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)"
Integration with g-Storage

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

![Diagram showing integration with g-Storage]
Outline

● Introducing GenomeCloud
  ● Who we are, why we start
● Galaxy on the GenomeCloud
● **Use cases and lessons learned**
● Conclusions
Use case

- Bioinformatics education support
Use case

- 4 half-day bioinformatics workshops
  - Date: 13\textsuperscript{th}, 20\textsuperscript{th} May, 3\textsuperscript{rd}, 10\textsuperscript{th} 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 enough for further research
- Lessons learned:
  - Be aware of bottlenecks
  - Fix it or adapt to it
Bottlenecks

- DBMS: fix by tuning configuration
- Get data: distribute
- Data transfer from shared storage: be aware and adapt
Network bottleneck

- Some NGS tools (e.g., BWA, bowtie ...) saturate the network bandwidth
- ✗ expect 5 to 10 minutes lead time
Lessons we learned

• Don’t re-invent the wheel
  • Similar demands and questions were asked and answered
  • 2 anecdotes (run_reports.sh, data manager)
• The more you know, the more you will find the value of Galaxy
Conclusions

Galaxy on the GenomeCloud provides more use cases to the Galaxy Community

- Geography
- Infrastructure
- Users

Galaxy is becoming a door to the diverse bioinformatics research in Korea.
And,
GenomeCloud helps them unlock the door.
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