Flexible Virtual Computing Platform on Clouds for Scientific Workflows

Kai Liu, Kento Aida, Shigetoshi Yokoyama, Yoshinobu Masatani
National Institute of Informatics

Abstract

Cloud platforms are expected to be promising solutions for conducting scientific computation, as they can provide powerful, elastic and economical computation ability. However, deploying a complete scientific computing platform on clouds is a challenging job even for experienced programmers, not to mention domain scientists who are lack of computer skills. In this poster, we propose a flexible virtual computing platform to run scientific workflows. Thanks to Docker[1], our proposed architecture enable users to easily deploy the scientific computing platform on clouds. Moreover, users have full flexibly to choose and combine all components, including scientific workflow systems, container cluster systems and storage systems.

Scientific Computing Platform

The scientific computing platform discussed in this poster consists of a scientific workflow system, a container cluster system and a storage system as illustrated in Figure 1.

Systems composing the target platform depend on plenty of libraries, binaries and configuration files; therefore, for users, **deploying or reproducing the scientific computing platform on clouds is a quite challenging job due to the system complexity.**

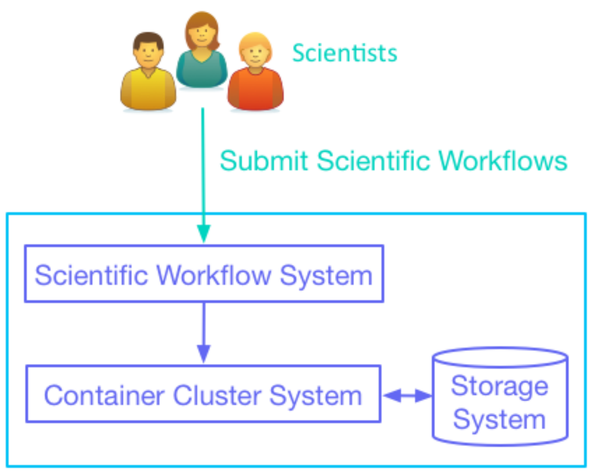

Figure 1: Scientific Computing Platform

Current Issues

Container Technology

By packaging software systems in Docker images, we can make software installation and configuration less painful. For example, Skyport[2] greatly reduces the heavy burden of deploying the computing platform, AWE/Shock, used in the biology community by utilizing Docker.

Limited Flexibility for Existing Workflow Systems

However, it has limitation of flexibility in the configuration of the computing platform. In the existing system, such as Skyport, a workflow system is combined with the proprietary cluster/storage systems. **The user cannot combine the workflow system with other cluster/storage systems.**

Proposed Virtual Computing Platform

Packaging Platform into Docker Containers

In our proposed architecture, the whole scientific computing platform, including a scientific workflow system, a container cluster system, a storage system as well as scientific tools (applications), run within Docker containers. Our middleware, Virtual Cloud Provider[3], can help users **automatically deploy the virtual computing platform on any cloud.**

Integrating Galaxy with Container Cluster Systems

Galaxy[4] is a widely used scientific workflow system in biology community. As showed in Figure 2, the scientific computing platform for running Galaxy workflows is consisted of Galaxy and three popular container cluster systems—Mesos/Aurora, Kubernetes, and Docker Swarm. By doing integration, **we can provide great flexibility for biologists to submit Galaxy tasks to their preferred container cluster systems.**

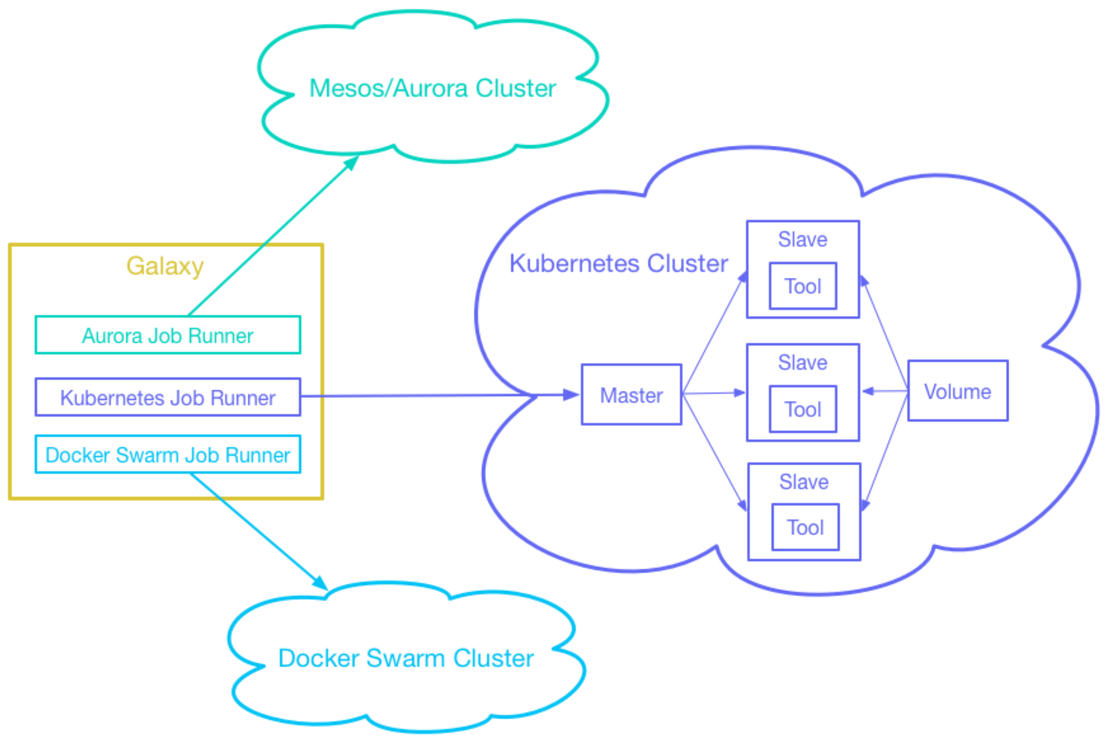

Figure 2: Scientific Computing Platform for Galaxy Workflows

- Galaxy provides a web interface for submitting workflows.
- Cluster master schedules tasks to cluster slaves.
- Cluster slaves execute Galaxy tools within it.
- Galaxy tool containers run within slave containers by using Docker-in-Docker mechanism.
- Data are stored within a Docker container, which called data volume container.

References

1. Docker, [Online]. Available: https://www.docker.com/. 2015.
2. W. Gerlach, et.al., “Skyport - Container-Based Execution Environment Management for Multi-cloud Scientific Workflows”, 2014 5th International Workshop on Data-Intensive Computing in the Clouds, 2014.
3. S. Yokoyama, et.al., "Middleware for Building Virtual Resources on Inter-Cloud", 2nd Annual Meeting on Advanced Computing System and Infrastructure, 2016.
4. Goecks, J. et.al., "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences”, Genome Biology, 2010.

National Institute of Informatics
{liukai, aida, yoko, ym}@nii.ac.jp