Kepler-XSEDE: A Scalable Workflow Framework

Shweta Purawat[1], Kevin Khuong[2], Parth Shah[2], Shava Smallen[1] and Ilkay Altintas[1]

[1]San Diego Supercomputer Center, University of California, San Diego, USA
[2]Department of Computer Science and Engineering, University of California, San Diego, USA

Abstract

The Kepler scientific workflow system [1] is an open-source, cross-project collaboration used in a wide variety of projects to manage, process, and analyze scientific data. Kepler provides a graphical user interface for designing scientific workflows. Kepler can be used to build Big Data workflows in diverse scientific domains [2]. Kepler supports Big Data applications using distributed-execution patterns like MapReduce and other Big Data programming patterns by running sub-workflows on engines like Hadoop, Spark, and Flink. Scientists from different disciplines have applied Kepler to create programmable, scalable, and reproducible workflows. Kepler-XSEDE initiative presented in this poster enables researchers to exploit the power of distributed XSEDE (Extreme Science and Engineering Discovery Environment) resources [3] using the composition, management and big data capabilities of Kepler workflow system.

The deployment of the Kepler workflow system on XSEDE resources will enable scientific users to develop and execute their scalable, programmable and reproducible Kepler workflows on XSEDE HPC clusters. It will eliminate the extra effort needed from XSEDE-Kepler users to download, install and configure Kepler on XSEDE; likewise, XSEDE admin or Kepler developer's staff-hours in assisting Kepler installation and configuration would not be required. Thus, keeping the complexity of execution of Kepler workflows on XSEDE resources to a minimum for scientific users. It will serve as a complete system to build, test and execute scalable workflows in high performance distributed environments.

Kepler will be available as an XSEDE Enterprise Service and on compute nodes of XSEDE SPs. The design option will enable users to start Kepler from a central VM login node at IU, `kepler.xsede.org`. The user accounts get created automatically, and when a user logs in, an X509 proxy credential will be set up so that the user has GSI-SSH capability to XSEDE login nodes. Users would be able to execute smaller workflows locally or leverage a GSI-SSH Kepler Actor to remotely execute a compute- and memory-intensive workflow on XSEDE resources. The design supports all Kepler workflow categories: Single-node Kepler workflows, HPC Kepler workflows, Big Data Kepler workflows and Heterogeneous workflows.

The initial version of the Kepler-XSEDE framework will enable easy and interactive command line execution of all types of Kepler workflows on the XSEDE resources. In the future, we will have a scientific gateway that will provide the capability to build and execute the Kepler workflows through a web-based user interface.

Keywords: Research infrastructures, Cyberinfrastructure, Scientific Workflows, Scientific computing, Kepler, XSEDE, Big data, distributed computing, HPC, Supercomputers

References

[1] Kepler: an extensible system for design and execution of scientific workflows. Altintas, I. and Berkley, C. and Jaeger, E. and Jones, M. and Ludascher, B. and Mock, S. (2004) Scientific and Statistical Database Management, 2004. Proceedings. 16th International Conference on. Pages: 423– 424.

[2] Jianwu Wang, Daniel Crawl, Ilkay Altintas, and Weizhong Li. Big Data Applications using Workflows for Data Parallel Computing. IEEE Computing in Science & Engineering, July-Aug 2014.

*Kevin Khuong and Parth Shah contributed equally to this work.

[3] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D. Peterson, Ralph Roskies, J. Ray Scott, Nancy Wilkins-Diehr, "XSEDE: Accelerating Scientific Discovery", Computing in Science & Engineering, vol.16, no. 5, pp.62-74, Sept.-Oct. 2014, doi:10.1109/MCSE.2014.80