Frameworks to Support Multi-Platform Distributed Data Intensive Applications

Shantenu Jha, Dan Katz and Jon Weissman
(LSU)

Abstract

Developing applications to work with large-scale distributed data is difficult. One approach to tackle challenges arising from such distributed data is the creation of data-intensive computing frameworks that can be utilized by multiple scientific applications, which provide support for several abstractions and which scale-out and are interoperable across multiple infrastructure. These are challenging, and often conflicting and constraining requirements. Some characteristic features of such scalable, extensible and infrastructure independent frameworks are that they support data-locality, affinity and multiple-patterns. In this talk we will introduce SAGA (http://saga.cct.lsu.edu) and show how it has been used to prototype and implement such frameworks. Specifically, we will discuss how SAGA has been used to provide (i) programmatic capability to manage distributed data functionality, and (ii) build the run-time environments and systems that help facilitate the requirements across distributed resources. We will discuss some preliminary results where SAGA has been used to support different patterns such as Map-Reduce and All-Pairs, over multiple heterogeneous infrastructure ranging from TeraGrid-like production systems to canonical Clouds (EC2, and the Open Cloud Testbed) concurrently.