In response to ASCR Program Announcement

Tigres: Template Interfaces for Agile Parallel Data-Intensive Science

Lead Principal Investigator
D. Agarwal Lawrence Berkeley National Laboratory daagarwal@lbl.gov 510-486-7078

Abstract. DOE User Facilities and large science collaborations are increasingly generating large enough datasets that it is no longer practical to download them to collaborator's institutions. They are instead stored at centralized compute and storage resources such as high performance computing (HPC) centers. Analysis of this data requires an ability to run on these facilities, but with current technologies, scaling an analysis to an HPC center and to a large dataset is difficult even for experts. This project is addressing the challenge of enabling collaborative analysis of DOE Science data through a new concept of reusable "templates" that enable scientists to easily compose, run, and manage collaborative computational tasks. These templates define common computation patterns used in analyzing a dataset.

Our approach is inspired by the success of the MapReduce model. When the MapReduce model emerged from the Internet search space, it provided a radically simpler paradigm for parallel analysis by certain classes of applications. The simplicity of the API and analysis model enabled many applications to quickly script powerful scalable analyses. We propose to follow the example of MapReduce and provide abstractions that support a wide-array of common scientific application computational patterns.

We will focus on four research topics in this proposed work. First, we will design and implement the template abstraction to capture the core set of fundamental workflow patterns, which will allow users to compose collaborative workflow scripts in a programming language of their choice. Second, we will design and develop a hybrid execution mechanism for the templates that enables users to prototype their analysis workflows on desktops and seamlessly adapt them to run in production environments at scale. Third, we will provide programmatic interfaces that will allow automated and user-provided provenance tracking. Finally, we will provide interfaces to capture execution state and allow users to understand complex parallel faults encountered during execution of the workflow.

Examples of DOE science areas where this capability will make a significant difference include light-source science, cosmology, combustion, atmospheric and soil carbon, next generation ecosystem experiments, materials simulation, and gene sequencing. We will engage with our collaborators in these domains in a user-centered design process to define the templates, execution, provenance, and fault-tolerance interfaces. The resulting capability is expected to dramatically expand the access to collaborative scientific data analysis.

The key outcome of this project will be the Tigres template library that will allow users to quickly develop and test their analysis workflows on laptops and desktops and then subsequently move them into production HPC resources and large data volumes. Tigres is expected to *significantly impact scientist productivity in collaborations by reducing complexity of composition, improving execution efficiency, and sharing of analysis templates in collaborations*.