An Integrated Middleware Framework to Enable Extreme Collaborative Science

Shantenu Jha (PI, Rutgers), Daniel S. Katz (Univ. of Chicago), Jon Weissman (Univ. of Minnesota)

Objectives: Extreme-scale collaborative science requires that the complexity of the underlying environment in terms of diverse storage, network, and computing platforms must be managed on the one hand, yet exploited on the other. The evolution of science applications in terms of new algorithms, improved fidelity, and integration of data, has proceeded in parallel with the evolution of advanced network services such as, for example, QoS on optical networks. However, the evolution of middleware that glues the two layers together has evolved much more slowly: lower-level services are often hard to use and represent a disruption to the flow of the application. Going the other direction, mechanisms to supply critical application characteristics that could shape the runtime behavior of the application with respect to low-level resource usage are not routinely available. A structured and standard approach to addressing these concerns does not exist, either at the middleware level, or in the form of services or tools. This leads to isolated, repeated, and non-extensible solutions and is not a scalable solution in the long run. What remains is a gap between resource capabilities and application requirements. It is the role of middleware to bridge this gap. But how should it be organized and presented to end-users and developers? What features must this middleware provide? How can semantic richness be balanced against feature creep and complexity?

Description: This project aims to bridge the gap between application requirements and diverse and heterogeneous platforms, by developing a middleware framework that can support the needs of tools and services in support of distributed scientific collaborative applications at extreme scales. We consider a variety of distributed applications, including those that operate on rich data pipelines in a distributed collaborative environment including: data generation and capture, data preprocessing, data analysis, and data storage and delivery. For these applications, tools and services are needed at all levels of this pipeline to enable data discovery, data transmission and streaming, data placement and storage, resource discovery, computation scheduling, and co-scheduling. Other types of applications share most of these needs.

A framework-based middleware provides an integrated way to address the many co-dependent issues in extreme-scale environments such as the emergence of disparate resource platforms and network capabilities, the inherent distribution of compute and data, and multiple-levels of application and run-time decision making. We propose a middleware framework that provides powerful abstractions for distributed computational and storage resources, and containers for computational tasks and distributed data. This project will address fundamental research challenges required to realize these abstractions, including techniques to enable resilience and performance for both data and computation. Our framework will enable a varied set of tools to be more easily constructed, such as workflow systems and in-situ data processing, to name a few.

Benefits and Outcome: This proposal has elements of research and development; it will deliver software solutions that will be usable by multiple DOE application science and tool developers. The project team will first validate the approach by building prototype tools using the framework building-blocks, deploy them on DOE and other systems available to the PIs, and apply them to real distributed collaborative DOE applications, such as those in the areas of Earth sciences and genomics.