Project Summary

The advent of extreme-scale collaborative sciences has demonstrated the potential for broad scientific communities to pool globally distributed resources to produce unprecedented data collections, simulations, visualizations, and analysis. System resources including supercomputers, data repositories, computing facilities, network infrastructures, storage systems, and display devices have been increasingly deployed around the globe. These resources are typically shared by large science communities over Internet or dedicated networks and hence exhibit an inherent dynamic nature in their availability, reliability, capacity, and stability. Scientific applications based on computational techniques with diverse physical, chemical, climatic, and biological models entail unique computing solutions with multifarious requirements. Application users, who are primarily domain experts, oftentimes need to manually configure their computing tasks over networks in an ad-hoc manner, which significantly limits the productivity of scientists and constrains the utilization of resources. The success of these extreme-scale distributed applications requires a massively scalable and highly adaptive distributed computing platform that provides optimized computing and networking services.

This proposal aims to develop a high-performance workflow system for extreme-scale scientific applications, whose primary goal is to enhance scientific productivity and facilitate research and education collaborations within a unified computing platform that is made highly accessible to scientists. The proposed system comprises a set of autonomous virtual nodes that optimize distributed computation-computing tasks over networks to meet various performance requirements in response to user interactions and time-varying network and host conditions. We will develop a number of analytical cost models to estimate data transport and processing time of specific visualization and general computing modules. We will investigate the computational complexity of mapping computing tasks as simple as linear pipelines or as complex as graph workflows onto heterogeneous networks and develop efficient mapping algorithms to optimize end-to-end application performance. We will delve into the stability and reliability issues of distributed computing workflows under various mapping and resource constraints. We will develop a collaborative metadata management mechanism that allows for accessing, managing and sharing metadata in a distributed environment. We will also develop a simulation-based formal method to test and evaluate component solutions before deploying the system in real network environments. The successful completion of this project will lay down a solid mathematical foundation for the analysis and control of system dynamics of large-scale scientific applications and produce a suite of critical computing and networking tools to support scientific collaborations.

Technical merits: (1) The proposed system automates the entire scientific discovery process of simulation, filtering, transport, visualization, and computational monitoring and steering over wide-area networks. (2) The low polynomial-time computational complexity of the workflow mapping algorithms ensures the scalability of the system in handling large network sizes and complex task structures. (3) The control strategy of adaptive reconfiguration incurs the least overhead for maximal network performance in coping with both user and resource dynamics. (4) The runtime performance analysis rigorously determines the stability of and maximizes the reliability of distributed computing workflow execution. (5) The simulation program allows us to quickly and accurately verify the validity of cost models, examine the correctness of optimization algorithms, and evaluate the runtime performance of the proposed methods in a simulated environment.

Broader impacts: (1) The proposed system will benefit a broad spectrum of DOE’s mission-critical scientific applications by providing scientists important capabilities to conduct remote computational steering with guaranteed performance, which will significantly improve their productivity by cutting down the computation-computing cycle time. (2) The resulted software will help the science community form larger collaborative teams and untangle them from constantly searching for and installing complex computing and networking tools so that they can concentrate on their original scientific goals. (3) The project activities will be directed toward promoting scientific education through creating opportunities for under- and graduate students, especially female and African American students to engage in research and developing new course curricula in modern computing and networking technologies.