SHF:Medium:Collaborative Research:
Compute on Data Path: Combating Data Movement in High-Performance Computing

Overview: The objective of this project is to develop a *Compute on Data Path* paradigm via an *integrated data-centric programming model, runtime system, data model, and storage system* to combat the increasingly critical data movement issue in high-performance computing (HPC). While the peak computational performance of HPC systems has crossed the petaflop mark and is moving to reach the exaflop range, the data movement and data access performance of HPC systems lag far behind. Along with big data applications such as information retrieval and data mining, HPC applications in critical areas of science and technology are also becoming increasingly data intensive. Data movement has become the killer bottleneck of modern HPC, for both performance and energy efficiency. Conventional HPC systems, however, are *computing-centric* and designed to utilize the computation capability of CPUs. In this project, the PIs propose a novel *Compute on Data Path* execution model to combat the data movement bottleneck in HPC. The new concept will be realized upon the PIs' successful developments in high-level programming models and optimizing compiler, data model, file systems, and I/O runtime. It is intended to deliver significantly better performance and energy efficiency than the existing execution models for data-intensive applications, and the solutions in this project will have a profound impact on algorithm design and system development in HPC as well.

Intellectual Merit: The *Compute on Data Path* paradigm is a new concept of *fusing data motion and computation* to minimize data movement and to deliver highly efficient HPC systems for data-intensive scientific discovery and innovation. By including computation capability on the data path, this methodology will significantly reduce data transfer to and from storage, network, and memory hierarchy of HPC systems, thus achieving compute on data in situ with high efficiency. The realization of the Compute on Data Path paradigm consists of four components: a data model that encapsulates and binds computations with data, a programming model that enables programmers to define and describe data motion and computation, a storage system that provides object-based storage for computation and data objects, and a runtime system that enables computations on the data path pipeline. The paradigm essentially provides a new execution model that shifts data-intensive operations and computation-intensive operations close to storage and compute nodes, respectively, and as much as possible on the data path. This *data-centric* execution model are fundermentally different than the traditional *computing-centric* model for HPC applications and systems. Considering the recent success of the MapReduce framework for data-intensive applications for enterprise computing, the Compute on Data Path paradigm can be a transformative innovation in the high-performance scientific computing domain and deserves an investment.

Broader Impacts: This multi-institution collaboration will integrate the development, education, and outreach efforts of several universities. Given the growing enormous need for supporting data-intensive sciences, the PIs will develop the Compute on Data Path HPC paradigm based on cutting-edge techniques in data-centric software, architectures, and platforms. The PIs will coordinate with institutional projects at Texas Tech University, the University of Houston, and Northwestern University to attract underrepresented students into this project. The experience will be integrated into undergraduate and graduate courses in majors such as computer science and computational science. The goal of the education plan is to elevate the emerging issue of supporting data-intensive HPC to a level commensurate with its increasing importance and to train a broadly inclusive and globally competitive science workforce by integrating research and development into the educational curriculum. Moreover, this project will provide a pathway to future national HPC systems to support data-intensive sciences, and it can have a direct impact on building exascale HPC machines. The proposed research will advance a broad range of fields that need HPC simulations for scientific discovery.

Keywords: data-intensive computing; big data; high-performance computing