

High-Performance Decoupling of Tightly Coupled Data Flows

The need to distill enormous amounts of data into useful knowledge is pushing the limits of computational science. In many cases, to achieve high performance, programmers tightly couple the data analysis with data generation -- making the analysis interdependent and closely coordinated with the computation. Often, however, tight coupling limits the flexibility provided by individual modules. To address this issue, this project will explore a hybrid approach that combines both types of coupling---tight and loose---in effect decoupling tightly coupled applications. The proposed work will result in a library of four data flow primitives: selection, aggregation, pipelining, and buffering. The research will include a software stack of three layers: a high-level data description of scientists' datasets, a data flow model built from the basic primitives, and a transport layer that will move data. A set of resilience strategies will cross-cut all three layers to enable the system to continue operating properly in the event of a failure of a component. The project will also produce a method for automatically constructing data flows from these primitives, designed as a generic solution that other workflow and coupling tools can use. The software will be evaluated in three science applications: fluid dynamics, superconductivity, and cosmology. The aim is to improve performance, reduce power, mitigate errors, and enhance usability.