

The Virtual Data Grid: A New Model and Architecture for Data-Intensive Collaboration

Ian Foster

Argonne National Laboratory
University of Chicago
foster@mcs.anl.gov

Abstract

It is increasingly common to encounter communities engaged in the collaborative analysis and transformation of large quantities of data over extended periods of time. I argue that these communities require a scalable system for managing, tracing, exploring and communicating the derivation and analysis of diverse data objects. Such a system could bring significant productivity increases facilitating discovery, understanding, assessment, and sharing of both data and transformation resources, as well as facilitating the productive use of distributed resources for computation, storage, and collaboration. I define a model and architecture for a virtual data grid capable of addressing these requirements. I define a broadly applicable model of a "typed dataset" as the unit of derivation tracking, and simple constructs for describing how datasets are derived from transformations and from other datasets. I also define mechanisms for integrating with, and adapting to, existing data management systems and transformation and analysis tools, as well as Grid mechanisms for distributed resource management and computation planning. Finally, I report on successful application results obtained with a prototype implementation called Chimera, involving challenging analyses of high-energy physics and astronomy data.