

Virtual Services in Data Grids

Arun Jagatheesan^{1,2}, Reagan Moore¹, Arcot Rajasekar¹ and Bing Zhu¹

¹San Diego Supercomputer Center, University of California at San Diego, La Jolla, CA 92093

²High Energy Experiment Group, University of Florida, Gainesville, FL 32611

{arun, moore, raja, bzhu}@sdsc.edu

Abstract

Data grids enable next generation scientific explorations that require intensive computation and analysis of petabyte-scale shared data collections. Apart from the challenge in creation and management of the data, another major challenge is the discovery of derived data products that have already been created. This work addresses the later challenge in minimizing response time and conserving the computing cycles in the grids by using “Virtual Services” to access derived data products. We analyze the role of virtual services for long running services in a service-oriented architecture and contribute a technique called “semantic mirror”. A prototype was developed to discover derived data products of grid services using SRB-MCAT technology and GriPhyN Virtual Data Language (VDL).

Introduction

One of the major challenges in grid-computing infrastructure is to manage huge amounts of data distributed over multiple sites. Derived data products are made as result of compute intensive scientific simulations or transformations. The cost of re-computing materialized derived data using multiple supercomputers and clusters is far more expensive than discovering the derived data product and moving it to the required site. Hence, before any long running service is offered in the grid to compute data, suitable mechanisms are required to find out if the same computation was already performed. If the same computation (with same input) has already been performed and the output from that computation is still available in the data grid, another compute intensive process need not be started. Thus the data grid can provide the existing result of a service to users without using the real service. This leads to the concept of *Virtual Services*.

Virtual Services in SOA.

Virtual Services provide the service without invoking the real service, thereby saving cost on computational resources and reducing the time of execution of related processes. In Data grids using Service

Oriented Architecture (SOA), service cachiers can be considered as the proxy of the real service provider. At a higher level of definition, the role of the service cachier is similar to dynamic web caching used in web servers. The difference is that the web caches store the objects that need to be sent along with the html pages, but the service cachier stores either the results of the computation or meta-data to point to a similar computational result without invoking the service.

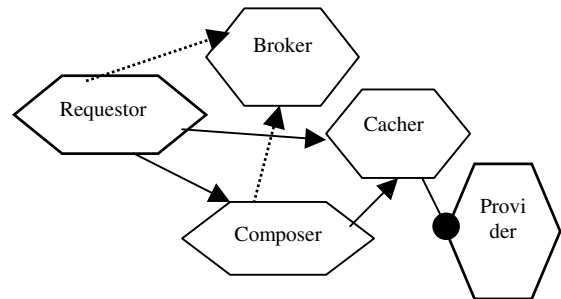


Figure 1. Service Invocation Roles

In semantic mirror technique for virtual service, we assume the existence of a *Grid Service Invocation Language* to invoke a grid service. In GriPhyN virtual data grid, the Virtual Data Language [1] acts as a novel “recipe maker” to describe computational procedures needed to materialize data that might not be available on the datagrid. In the prototype built, we used MySRB [2] to store the VDL documents that describe each invocation on the grid. Service Cacher provided virtual service using the SRB-MCAT [2] to query the data grid for the location of semantically equivalent derived data products that might be available before invoking the real service.

More analysis on the performance and integration with a new version of the VDL need to be done before being used in production systems.

[1] Foster, I., Voeckler, J., Wilde, M. and Zhao, Y., Chimera: A Virtual Data System for Representing, Querying, and Automating Data Derivation. In *Scientific and Statistical Database Management*, (2002).

[2] Rajasekar, A., Wan, M. and Moore, R., MySRB & SRB – Components of a Data Grid. In *High Performance Distributed Computing*, (2002).