Servicing Seismic and Oil Reservoir Simulation Data Through Grid Data Services*

Sivaramakrishnan Narayanan, Tahsin Kurc, Umit Catalyurek, and Joel Saltz

Department of Biomedical Informatics,
The Ohio State University, Columbus, OH, 43210
{krishnan, kurc, umit, jsaltz}@bmi.osu.edu

Abstract. This paper presents the implementation of a two layer infrastructure for servicing queries against large datasets generated in oil reservoir simulation studies in the Grid. The first layer implements *object-relational virtualization* of file-based dataset stored on a storage cluster. The second layer provides an implementation of Grid Data Services via Open Grid Services Architecture Data Access and Integration (OGSA-DAI) middleware.

1 Introduction

In an increasing number of engineering and science fields, the volume of data generated and processed is in the order of terabytes. Simulation-based oil reservoir management studies are an example of applications that generate and reference large volumes of simulation and experimental data. The objective is to develop complex numerical models of subsurface reservoirs and use these models to efficiently search for alternative oil production strategies in order to optimize profits and minimize adverse effects to the environment [20,28]. In this optimization process, there is a need to provide support for management and querying of large volumes of data, generated by simulations or collected from field measurements, in order to be able to refine model parameters and determine the next set of simulations to be carried out. In addition, the datasets can be generated and stored at multiple locations, since the computational requirements of the simulations may require use of machines at supercomputing centers.

There has been considerable progress in Grid computing technologies in recent years. In addition to a wide array of middleware systems and tools, a services-based view of the Grid has emerged. In this view, data sources and applications are exposed to the environment using standard interfaces. Users interact with the resources through well-defined Grid services protocols. In this way, the complexities and heterogeneity of individual resources can be hidden from clients and greater interoperability among applications can be achieved.

* This research was supported in part by the National Science Foundation under Grants #ACI-9619020 (UC Subcontract #10152408), #EIA-0121177, #ACI-0203846, #ACI-0130437, #ANI-0330612, #ACI-9982087, #CCF-0342615, #CNS-0406386, #CNS-0426241, Lawrence Livermore National Laboratory under Grant #B517095 (UC Subcontract #10184497), NIH NIBIB BISTI #P20EB000591, Ohio Board of Regents BRTTC #BRTT02-0003.

J.-M. Pierson (Ed.): VLDB DMG 2005, LNCS 3836, pp. 129–142, 2005.
© Springer-Verlag Berlin Heidelberg 2005

Several core functions need to be supported in an end-to-end system for enabling data-driven scientific applications in a Grid environment. These functions include management of data types and metadata, virtualization of data sources and data subsetting, data product generation (e.g., data aggregates from data subsets), and Grid services interfaces. In our work, we develop an integrated suite of middleware components to support these functions. These middleware components are shown in Figure 1. In this suite, DataCutter, which is a component-based middleware, enables combined use of task- and data-parallelism and is used to support data product generation (e.g., aggregates of data subsets) [7]; STORM [22,23] provides virtualization of file based datasets as object-relational tables and support for data subsetting; Mobius [19] supports management of data definitions and data types as XML schemas, XML virtualization of data, and metadata management. In an ongoing project, we are integrating these middleware systems with the Open Grid Services Architecture Data Access and Integration (OGSA-DAI) middleware toolkit [24] to allow access to the functionality provided by these components via OGSA-DAI Grid services protocols.

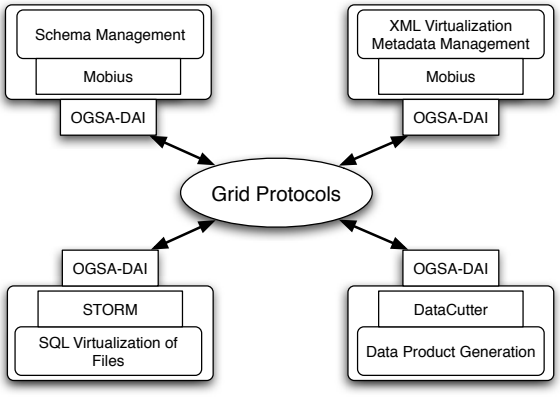

Fig. 1. Middleware components and toolkits to support data-driven scientific applications in the Grid

In this paper, we describe the design and implementation of a layered infrastructure for serving large, distributed datasets generated in oil reservoir simulation studies in a Grid environment using STORM and OGSA-DAI. The first layer in our infrastructure implements support for efficient use of distributed storage clusters and enables *object-relational virtualization* of file-based datasets. This layer builds on the STORM middleware framework. The second layer leverages the existing work in the Grid community to provide integrated access to datasets served by multiple STORM instances. This layer is implemented using the OGSA-DAI middleware toolkit. We describe the integration of STORM as a data source in OGSA-DAI and present a preliminary performance evaluation of the integrated system.