Chapter 2.11
An Architectural Overview of the GRelC Data Access Service

Sandro Fiore
University of Salento & CMCC, Italy

Massimo Cafaro
University of Salento & CMCC, Italy

Alessandro Negro
CMCC, Italy

Giovanni Aloisio
University of Salento & CMCC, Italy

Salvatore Vadacca
CMCC, Italy

Roberto Barbera
Università di Catania and Istituto Nazionale di Fisica Nucleare, Italy

Emidio Giorgio
INFN Sez. di Catania, Italy

ABSTRACT

Grid computing is an emerging and enabling technology allowing organizations to easily share, integrate and manage resources in a distributed environment. Computational Grid allows running millions of jobs in parallel, but the huge amount of generated data has caused another interesting problem: the management (classification, storage, discovery etc.) of distributed data, i.e., a Data Grid specific issue. In the last decade, many efforts concerning the management of data (grid-storage services, metadata services, grid-database access and integration services, etc.) identify data management as a real challenge for the next generation petascale grid environments. This work provides an architectural overview of the GRelC DAS, a grid database access service developed in the context of the GRelC Project and currently used for production/tutorial activities both in gLite and Globus based grid environments.

INTRODUCTION

Grid computing (Berman, 2003) is an emerging and enabling technology allowing organizations to easily share, integrate and manage resources in a distributed environment. As an advanced form of distributed computing, grids link together servers, data sources, sensors and applications into a single system by means of a specific glue named *grid middleware*. All of these components can be very heterogeneous (different operating systems, multiple hardware platforms and architectures) and geographically dispersed.

Starting from the first distributed testbeds (I-WAY, GUSTO, etc.) a lot of efforts have been devoted to the improvement of grid services to

DOI: 10.4018/978-1-4666-0879-5.ch2.11

support enhanced job submission, efficient file transfer, distributed resources monitoring and so on. However, the early attempts of grids were mainly related to the aggregation of computational power trying to address large scale computational problems. These efforts demonstrated the potential of a Computational Grid to run millions of jobs in parallel, but the huge amount of generated data has caused another interesting problem: the management (classification, storage, discovery etc.) of distributed data, i.e., a Data Grid specific issue.

In the last decade many efforts concerning the management of data (grid-storage services, metadata services, grid-database access and integration services etc.) identify data management as a real challenge for the next generation petascale grid environments. Raw data management, that relates to storage services, file transfer protocols, reliable file transfer services, storage resource managers etc. are obviously very important, but may be useless without something enabling (i) the discovery of a file within a distributed collection, (ii) metadata search and browsing, (iii) the classification of a set of output results, etc. These features make feasible the management of such a high scale production activity.

Both system (to support grid services) and application-level (for researchers and scientists) metadata often rely on relational and XML databases that can obviously be distributed and heterogeneous. Grid Services for database access and integration (Watson, 2003) are now capturing the interest of the grid community since they play a strategic role and provide added value to a grid production environment.

The biggest European Production Grid (EGEE) aims at providing researchers in academia and industry with access to major computing resources. The EGEE infrastructure, composed of standard PCs interconnected through high performance links on the Internet, is a suitable infrastructure for handling a large number of computational tasks. Research and development activities related to EGEE have generally focused on cpu-bound applications where data is stored in files.

Nowadays, in the EGEE project, a special attention has been paid to grid database access because there is an urgent need to interconnect pre-existing and independently operating databases. This requirement, addressed by the GRelC middleware, is common to many e-Science applications willing to use the EGEE grid environment; in particular, among the others, to the bioinformatics and astrophysics communities.

BACKGROUND

This Section introduces the Grid Database Access Service concept, describing the main existing approaches (front-end and embedded) and then dealing with two case studies: GRelC and OGSA-DAI. We will also highlight the main differences between the two research projects from several points of view: programming language, client support, security, data access and integration services, etc. Finally, we will briefly discuss the convergence issues related to the novel OGF DAIS specifications.

Grid Database Access Service: General Description

A Grid Database Access Service enables the virtualization of both relational and non- relational (i.e., XML-DBs) database systems within a Grid environment. It must provide secure, transparent, robust and efficient access to heterogeneous and distributed databases exposing standard interfaces to enable interoperability with other grid components and/or services.

Several research projects exploit the *service-in-the-middle* or *front-end* approach to provide such kind of functionalities, that is, they focus on the development of a transparent, secure and robust grid interface to existing DBMSs.

Related Content