Enabling Access to Astronomical Databases through the Grid: a Case Study

S. Pastore ⁽¹⁾, A. Volpato ⁽¹⁾, A. Baruffolo ⁽¹⁾, L. Benacchio ⁽¹⁾,
G. Taffoni ⁽²⁾, R. Smareglia ⁽²⁾, C. Vuerli ⁽²⁾, R. Crevatin ⁽²⁾

⁽¹⁾ INAF – Osservatorio Astronomico di Padova, ⁽²⁾ INAF – Osservatorio Astronomico di Trieste

Abstract

In recent years great efforts have been spent in the astronomical community to allow a global and seamless electronic access to distributed astronomical data repositories, and to enable scientific analysis on them. A system capable of providing these services over distributed databases and computer resources is called a Virtual Observatory (VO). VOs are still in the study phase, with many projects around the world now delivering their first prototype implementations. In the meanwhile grid technologies have started to emerge and consolidate and are now expected to play a fundamental role in the development of Astrophysical Virtual Observatories. We report here about our activities aimed at integrating in the grid environment a system, developed in Padova, specifically designed for accessing very large astronomical catalogues. We evaluated several possible solutions, including the use of a tool for accessing databases developed by the European Data Grid (EDG) project, but finally we decided to adopt a web services based architecture, retaining the security infrastructure provided by the EDG software.

We plan to evolve towards an implementation fully compliant with the latest Open Grid Service Architecture specification, but without losing compatibility with the grid middleware adopted in our project. This work is being conducted in the framework of Grid.it, a project, funded by the Italian Ministry for Education, University and Research, aimed at studying a grid infrastructure for scientific research in Italy and developing specific software tools and prototype applications that run on it.

This catalogue access system will eventually become one of the first building blocks of the Italian Virtual Observatory, currently under development in the framework of a closely-related project, named DRACO. To this end, starting from the design phase of our system, and where applicable, we adopted the standards set forth by the International Virtual Observatory Alliance (IVOA).

The framework

The Astronomical Observatory of Padova (OAPd) Catalog group is involved in the Grid.it project, a multidisciplinary research project, funded by the Italian Ministry for Education, University and Research, aiming at "Enabling platforms for high-performance computational grids oriented to scalable virtual organizations". The Grid.it project is organized around fifteen Work Packages (WPs) focused on implementing a grid infrastructure on a national scale (based upon the GARR scientific network), creating a grid-specific set of software tools and on developing some demonstrators within several applicative fields, ranging from Computational Chemistry to Earth Observation.

WP10 (Grid Applications for Astrophysics) of this project collects the various groups whose main goal is exploring the use of grid technologies for the development of astrophysical applications. Within this work package the main goal of the OAPd group is to study the portability to the Grid of an existing system for the consultation of large astronomical catalogues, currently serving on the net the second Guide Star Catalog (GSC-II).

In the same framework, OAPd closely collaborates with the Technology Group at the Astronomical Observatory of Trieste (OAT), which is dealing with the similar problem of integrating in the Grid the archive of observational data from the Italian Galileo National Telescope (TNG).

The research field is constituted by Grid technologies: the set of hardware, protocols and software packages that enable coordination and sharing of highly distributed computational and storage resources (fig.1). Such environment supports, efficiently and in a secure and controlled way, resource sharing and integration in a scalable Virtual Organization (VO) context. VOs (sets of individuals or organizations that need to share resources to solve a given complex problem) have a strong dynamic structure, form, lasting and composition.

The testbed for grid applications is the Italian INFN Production Grid for Scientific Applications (<http://grid-it.cnaf.infn.it>).

Fig. 1: layers in the Grid.it infrastructure

INFN-Grid middleware release 2.0.0 is derived from LCG (LHC Computing Grid) distribution, provided by the grid deployment project for High Physics applications related to the Large Hadron Collider. LCG itself is a stable version of the middleware released by EDG (European Data Grid) project.

At present time, the production grid supports 18 sites and 13 Virtual Organization.

Our Case Study

Goal of the project is to integrate into the grid a system for accessing large astronomical catalogues. The existing system (fig.2) allows querying collections of catalogues by means of a web interface. The catalogue data and metadata are managed by an ORDBMS Object-relational database Management System (Informix Dynamic Server) enhanced by the PosAstro DataBlade, a module specifically developed for handling spatial information for astronomical objects that provides support for astronomical data types and an R-Tree based index to optimize query execution. The server side is implemented by a pool of java servlets; the use of CORBA technology allows interaction with legacy software.

Fig.2: existing GSC2 Access System

Design-time considerations

We decided to implement our prototype following the web services paradigm because:

- this is one of the most stable and widely used architectures for building distributed applications;
- the Grid reference architecture (Open Grid Services Architecture, OGSA) is based in turn on a web services extension;
- although the Grid middleware of INFN production Grid doesn't support services at present time, they will likely be included by the end of 2004;
- currently it is under discussion within the Global Grid Forum (GGF) how OGSA will evolve in the near future, but since it will be most likely based on web services technology, our choice will not prevent us from following the evolution of architectural standards;
- we want to be compliant with International Virtual Observatory Alliance (IVOA) standards: IVOA recommends the adoption of web services and XML for implementing data source access;
- web services can be easily used as building blocks of more complex applications;
- it is straightforward to develop user interfaces (e.g. by the means of web browser and JSP technology);
- they can easily be integrated into existing Grid Portals.

System architecture

The system is composed by two web services: a Metadata Access service and a Query service (Fig. 3). The interaction between a client and the services is carried out by SOAP messages over https. Currently, the services can be accessed either interactively or in batch mode. Interactive access is performed through a JSP page that can be reached by a web browser. Batch access is carried out requesting the execution of a command line client application by submitting a Job Description Language (JDL) file from a machine with the User Interface Grid software installed. In both cases, the user must have a valid User Certificate installed on his machine.

Fig. 3 : current system architecture

The services

They are implemented in Java and both communicate via JDBC with the DBMS.

- **Metadata Access Service:** retrieves metadata information about catalogues hosted by the system by connecting with a 'lightweight' database running on the same machine as the service container. This service answers to two possible requests: to the standard web services request *?wsdl* with the WSDL file describing the service itself, and to the *info* request with an XML file containing the catalogue metadata in an IVOA compliant format.
- **Query Service:** this service answers to queries expressed in SQL or ADQL with a VOTable containing the result set. Currently, for test purposes, queries are performed on a MySQL database containing a subset of the data hosted by the full system. It answers also to the standard web services request *?wsdl* returning the WSDL file describing the service itself.

Security

Access to services is controlled following Grid Security Infrastructure (GSI) standard, based on the usage of X509 certificates both for Grid users and for Grid resources. We adopted the implementation provided by EDG Security: this package contains a modified version of Apache Tomcat (web application server) and Axis Engine (service container) along with the libraries needed to interact with User Certificates and the Grid middleware devoted to management of Virtual Organization membership. The EDG Security components are the Trust Manager and the Authorization Framework.

Security = Authentication + Authorization

The authentication of grid clients is performed by the Trust Manager on X.509 certificates basis.

The Trust Manager is a pure Java-based component for validation of X.509 certificate paths used in SSL/TLS connections to secure web services. It is integrated at server side in the Tomcat servlet container, substitutes and extends the default SSLServerSocketFactory.

The authorization is performed by the Authorization Framework, which makes the mapping from a grid client to an access role through the extraction of the subject Distinguished Name from the client certificate.

The Authorization Framework consists of the Authorization Manager and a wrapper to integrate authorization functionality into the web services engine.

In the request flow in front of the actual SOAP endpoint is put an AXIS handler that extracts the client certificate from the SOAP request, passes the subject DN and any optional interrogation parameters to the Authorization Manager, and stores the returned result in the MessageContext, which can then be retrieved by the protected service (Fig. 4).

Fig. 4 : interaction with EDG Security components

Deployment

Both services are installed on a computational node (Worker Node) connected to the Production Grid; they are deployed within Apache Tomcat web application server equipped with Axis, and secured by EDG security (Fig. 5).

Fig. 5. : system deployment and interaction with environment

Work in progress

In the current Grid architecture Information Services (discovery and monitoring of resources) are structured hierarchically. They encompass semantic and interaction models for computational (CE) and storage (SE) Grid resources. The corresponding models for data sources and services accessing them have not been defined yet.

In close collaboration with CNAF the architecture of the Italian Production Grid is being extended in order to deal with this limitation. In this way it will be possible to publish and discover our data sources and services, achieving a true integration with the Grid infrastructure (Fig. 6).

This process is being carried out in such a way that the description of data sources and services provided by the extended Grid Information Services will be allowed to be IVOA compliant whenever needed.

The task of extending the Italian Grid middleware is being carried out at CNAF, while the Information Provider software for the newly defined grid resources is being developed at the Astronomical Observatory of Padova.

Fig. 6 : integration with the Grid Information Services

Further developments

The next step will exploit the true integration of the newly defined resources into the Grid to provide additional functionalities; it will then be possible for an application:

- to save the output of a query on a Grid Storage Element either specified by the user or automatically chosen by the Resource Broker
- to replicate the output file and to register the copies (notifying its presence to the Grid) in the most convenient locations for further processing

Bibliography

- Albrecht, M, et al. 1996, "Astronomical Server URL", <http://vizier.u-strasbg.fr/doc/asu.html>
- Baruffolo, A. & Benacchio, L. 1998a, in ASP Conf. Ser., Vol. 145, Astronomical Data Analysis Software and Systems VII, ed. R. Albrecht, R. N. Hook, & H. A. Bushouse (San Francisco: ASP), 382
- Baruffolo, A. & Benacchio, L. 1998b, SPIE Proc., 3349, 274
- Baruffolo, A., Benacchio, L. & Benfante, L. 1999, in ASP Conf. Ser., Vol. 172, Astronomical Data Analysis Software and Systems VIII, ed. David M. Mehringer, Raymond L. Plante, & Douglas A. Roberts (San Francisco: ASP), 237
- Benfante, L., Volpato, A., Baruffolo, A. & Benacchio, L. 2001, in ASP Conf. Ser., Vol. 238, The OaPd System for Web Access to Large Astronomical Catalogues, ed. F. R. Harnden Jr., F. A. Primini, and H. E. Payne
- Documentation for EDG Security, <http://edg-wp2.web.cern.ch/edg-wp2/security/edg-java-security.html>
- Documentation for OGSA/DAI release 3, <http://www.ogsadai.org.uk/docs/R3/index.php>
- Documentation for release 2.1.7 of EDG-Wp2 Spitfire, <http://edg-wp2.web.cern.ch/edg-wp2/spitfire/index.html>
- The International Virtual Observatory Alliance, <http://www.ivoa.net>
- GRID.IT: An Italian National Research Council Project on Grid Computing Funded under the National Programme "*Fondo per gli Investimenti della Ricerca di Base (FIRB)* (2002-2005)", <http://www.grid.it>
- The INFN Production Grid for Scientific Applications, <http://grid-it.cnaf.infn.it>
- Foster I., Kesselman C., Tuecke S. The Anatomy of the Grid: Enabling Scalable Virtual Organizations, 2001
- P.Quinn, P.Benvenuti, P.Diamond, F.Genova, A.Lawrence, Y.Mellier, 2003 , "The Astrophysical Virtual Observatory (AVO): a Progress Report", Proc. SPIE 4846, 58.
- IVOA Architecture Overview 0.2, <http://www.ivoa.net/forum/architecture/0012.htm>
- Resource Metadata for the Virtual Observatory, IVOA Proposed Recommendation, 23/03/2004, <http://www.ivoa.net/Documents/PR/ResMetaData/RM-20040323.html>
- IVOA Identifiers, IVOA Proposed Recommendation, 31/10/2003, <http://www.ivoa.net/Documents/PR/Identifiers/Identifiers-20031031.html>
- The IVOA Data Modelling Working group, <http://www.ivoa.net/forum/dm/>
- P.Quinn, P.Benvenuti, P.Diamond, F.Genova, A.Lawrence, Y.Mellier, 2003 , "The Astrophysical Virtual Observatory (AVO): a Progress Report", Proc. SPIE 4846, 58.
- Roll, J. 1996, in ASP Conf. Ser., Vol. 101, Astronomical Data Analysis Software and Systems V, ed. G. H. Jacoby & J. Barnes (San Francisco: ASP), 536
- Space Telescope Science Institute, (STScI), Osservatorio Astronomico di Torino, 2001, "The Guide Star Catalog", Version 2.2 (GSC2.2) (2001), VizieR On-line Data Catalog: I/271