



Contribution ID: 68

Type: Oral contribution

Diligent and OpenDLib: long and short term exploitation of a gLite Grid Infrastructure

Wednesday 1 March 2006 14:15 (15 minutes)

The demand for Digital Libraries has recently grown considerably, DLs are perceived as a necessary instrument to support communication and collaboration among the members of communities of interest; many application domains require DL services, e.g. e-Health, e-Learning, e- Government, and many of the organizations that demand a DL are small, distributed, and dynamic, because they use the DL to support temporary activities such as courses, exhibitions, projects, etc.

Nowadays the construction and management of a DL requires high investments and specialized personnel because the content production is very expensive and multimedia handling requires high computational resources. The effect are that years are spent in designing and setting up a DL and that the DL systems lack interoperability and the services provided are difficult to reuse.

This development model is not suitable to satisfy the demand of many organizations, so the purpose of DILIGENT is to create a Digital Library Infrastructure that will allow members of dynamic virtual research organizations to create on-demand transient digital libraries based on shared computing, storage, multimedia, multi-type content, and application resources. Following this vision Digital libraries are not ends in themselves; rather they are enabling technologies for digital asset management, electronic commerce, electronic publishing, teaching and learning, and other activities.

DILIGENT is a is a three-year European funded project that aims at developing a test-bed DL infrastructure able to create a multitude of DLs on-demand, manage the resources of a DL (possibly provided by multiple organizations), and operate the DL during its lifetime. These DLs created by DILIGENT will be active on the same set of shared resources: content sources (i.e. repositories of information searchable and accessible), services (i.e. software tools, that implement a specific functionality and whose descriptions, interfaces and bindings are defined and publicly available) and hosting nodes (i.e. networked entities that offer computing and storage capabilities and supply an environment for hosting content sources and services).

By exploiting appropriate mechanisms provided by the DL infrastructure, producer organizations register their resources and provide a description of them. The infrastructure manages the registered resources by supporting their discovering, reservation, monitoring and by implementing a number of functionalities that aim at supporting the required controlled sharing and quality of service.

The composition of a DL is dynamic since the services of the infrastructure continuously monitor the status of the DL resources and, if necessary, change the components of the DL in order to offer the best quality of service. By relying on the shared resources many DLs, serving different communities, can be created and modified on-the-fly, without big investments and changes in the organizations that set them up.

The DILIGENT infrastructure is being constructed by implementing a service oriented architecture in a Grid framework. The DILIGENT design will be service oriented in order to provide as many reusable components as possible for other e-applications that could be created on top of the basic DILIGENT infrastructure. Furthermore, DILIGENT exploits the Grid middleware, gLite, and the Grid production infrastructure released by the Enabling Grid for E-Science in Europe (EGEE) project. By merging a service-oriented approach with a Grid technology we can exploit the advantages of both. In particular, the Grid provides a framework where a good control of the shared resources is possible. By taking full advantage of the scalable, secure, and reliable Grid infrastructure each DL service will provide an enhanced functionality with respect the equivalent non-Grid-aware service. Moreover, the gLite Grid enables the execution of very computational demanding applications, such as those required to process multimedia content. DILIGENT will enhance existing Grid services with the functionality needed to support the complex services interactions required to build, operate and maintain transient virtual digital libraries.

In order to support the services of the DILIGENT framework and the user community expectations some key Grid services are needed: the Grid infrastructure should support a cost-effective DL operational model based on transient, flexible, coordinated "sharing of resources", address the main DL architecture requirements (distribution, openness, interoperability, scalability, controlled sharing, availability, security, quality), provide a basic common infrastructure for serving several different application domains and offer high storage and computing capabilities that enable the provision of powerful functionality on multimedia content e.g. images and videos.

From the conceptual point of view the services that implement the DILIGENT infrastructure are organized in a layered architecture.

The top layer, i.e. the Presentation layer, is user-oriented. It supports the automatic generation of user-community specific portals, providing personalized access to the DLs.

The Workflows layer contains services that make it possible to design and verify the specification of workflows, as well as services ensuring their reliable execution and optimization. Thanks to these set of services it is possible to expand the infrastructure with new and complex services capable to satisfy unpredicted user needs.

The DL Components layer contains the services that provide the DL functionalities. Key functionalities provided by this area are: management of metadata; automatically translation for achieving metadata interoperability among disparate and heterogeneous content sources; content security through encryption and watermarking; archive distribution and virtualization; distributed search, access, and discovery; annotation; cooperative work through distributed workspace management.

The services of the lower architectural layer, the Collective Layer, jointly with those provided by the gLite Grid middleware released by the EGEE project, manage the resources and applications needed to run DLs. The set of resources and the sharing rules are complex since multiple transient DLs are created on-demand and are activated simultaneously on these resources.

Following the first tests performed on the first releases of the gLite middleware the following Grid requirements were identified: it should be possible to query for the maximum number of CPUs concurrently available in order to allow to a DILIGENT high level service to automatically prepare a DAG where each node will be entitled to process a partition of the data collection, to use parametric jobs/automatic partitioning on data, to support service certificate for a high level service, to specify a job specific priority, to specify a priority for a user or for a service, to ask for on-disk encryption of data, to dynamically manage VO creation and to dynamically support user/service affiliation to a VO.

DILIGENT will be demonstrated and validated by two complementary real-life application scenarios: one from the culture heritage domain, one from the environmental e-Science domain. The former is an interesting challenge thanks to

the multidisciplinary collaborative research, the image based retrieval, the semantic analysis of images, and the support for research and teaching. The latter obliges DILIGENT to manage a wide variety of content types (maps, satellite images, etc.) with very large, dynamic data sets in order to support community events, report generation, disaster recovery.

The DILIGENT project collaborates with EGEE mainly through technical interactions (technical meetings (mainly with JRA1), gLite mailing lists subscription, tutorial) and feedback on EGEE activities and on DILIGENT project (gLite bugs submission and grid related DL requirements).

Now DILIGENT has two independent infrastructures (gLite v1.4): a Development Infrastructure (DDI) and a Testing infrastructure (DTI). These infrastructures are geographically distributed, linking 6 sites in Athens, Budapest, Darmstadt, Pisa, Innsbruck and Rome. We are running gLite experimentation tests on these infrastructures since July 2005 and we collected some useful data about data and job management.

As first approach to exploit the gLite Grid storing and processing on demand capabilities, we developed two experimental brokers that, starting from an existing digital library management system, named OpenDLib, allow interfacing the DDI. The gLite SE broker provides OpenDLib services with the pool of SEs available via the gLite software. Moreover, it optimizes the usage of the available SEs. In particular, this service interfaces the gLite I/O server to perform the storage (put) and withdrawal (rm) of files and the access to them (get). In designing this service one of our main goals was to provide a workaround to two main problems, i.e. inconsistency between catalog and storage resource management systems, and failure without notification in the access or remove operations. Although the gLite SE broker could not improve the reliability of the requested operations we designed it in such a way to: (i) monitor its requests, (ii) verify the status of the resources after the processing of the operations, (iii) repeat the registration in the catalog and/or storage of the file until it is considered correct or unrecoverable, (iv) return a valid message reporting the exit status of the operation.

The gLite WMS wrapper provides to the other OpenDLib services with the computing power supplied by gLite CEs. Actually, the goal of this service is to provide an higher level interface than those provided by the gLite components for managing jobs, i.e. applications that can run on CEs, and DAGs, i.e. direct acyclic graphs of dependent jobs. The gLite WMS broker has therefore been designed to: (i) deal with more than one WMS, (ii) monitor the quality of service provided by these WMSs by analyzing the number of managed jobs and the average time of their execution, and, finally, (iii) monitor the status of each submitted job querying the Logging and Bookkeeping (LB) service.

Author: Dr PAGANO, Pasquale (CNR-ISTI)

Co-authors: Dr MANZI, Andrea (CNR-ISTI); Dr BERNARDINI, Davide (CNR-ISTI)

Presenter: Dr BERNARDINI, Davide (CNR-ISTI)

Session Classification: 1c: Earth Observation - Archaeology - Digital Library

Track Classification: Earth Observation - Archaeology - Digital Library