Introduction

The complexity of Grid security solutions and intricate access to core Grid services still presents an obstacle on the path towards wider adoption of computing Grids by scientific communities. Our work aims to provide efficient and easy access to data in the EGEE Grid from the GridSpace Virtual Laboratory and to minimize the learning curve involved in accesing LCG File Catalogues (LFC), storage elements and the Grid Security Infrastructure (GSI) by concealing most technical details.

The GridSpace Virtual Laboratory is a set of integrated components which, when used together, form a distributed and collaborative space for science [1]. With its many applications, including distributed decision support for infectious disease treatment [2], integrated access to domain databases, Grid–enabled access to media content [3] and integration of computational chemistry applications with the Grid infrastructure [4], it has been selected as one of the enabling platforms for the PL-Grid national project [5]. The GridSpace Virtual Laboratory is intended to run on top of any Grid. As EGEE (http://public.eu-egee.org/) is the largest production Grid infrastructure in Europe, a natural requirement is to provide access to EGEE storage and catalogue services, particularly the LCG File Catalogue [6]. The importance and advantages of this functionality cannot be underestimated since EGEE hosts immense amounts of data produced by various Grid projects. An example would be the Worldwide LHC Computing Grid (WLCG) collaboration, which, after years of preparation and numerous
simulations performed using Monte Carlo data, has recently started to store real data in EGEE and several other Grid systems [7].

In addition to solving the above mentioned problem, we have developed an innovative Data Source Registry for storing user credentials and data source parameters. Data access supported by Data Source Registry is a new approach that simplifies the complexity of contemporary Grid projects.

Objectives

The article presents the design of LFC DS -- a data access solution developed for GridSpace. The paper also presents the state of the art, conceptual view of the solution, file and directory management APIs, security measures, data registration processes and test results. In particular, it presents alternative virtual laboratories available for the computational science community as well as other projects that attempt to simplifyintegrate or expose databases or grid data access technologies as services. Furthermore, it demonstrates the diversity with which grid project approach data access and lists libraries providing access to gLite data resources.

Technical sections show an overview of our solution and role of each
of its components, explain how data access was implemented and delineate the design of file and directory management APIs. In addition, security aspects are considered, i.e. how multiple users can employ gLite libraries intended only for single-user utilization, by sharing credentials and encryption. A separate section covers mechanisms for data source registration, including user interface and backend mechanisms. The technical part of our paper ends with a presentation of performance evaluation results.

Finally, the summary section states the outcomes of our work and identifies its main applications.

Methodology used

We have devised an API for users of the GridSpace Virtual Laboratory that creates an abstraction of working with local files with no intervening GSI, i.e. with no Grid certificate-related operations (although the user works with files stored on the Grid with all GSI mechanisms in place).

The main contribution of this work is an easy and innovative method of programmatically accessing EGEE storage services that conceals technical details from scientific developers. The ease of access to data is often neglected when constructing APIs. In our approach, it forms a principal requirement. Moreover, the API we have created expects much less information from the user and, in addition, it integrates with the Virtual Laboratory environment which removes the requirement of executing end-user programs on gLite User Interface machines directly. Instead, they are run by the GridSpace Engine which handles all underlying grid data operations. The method we use to provide such functionality is original: we combine a Data Source Registry (DSR) with a custom-developed server utilizing gLite libraries. Furthermore, our solution is an interesting application of innovative Java libraries for distributed computing: Cajo and RMIIO. Although the Data Source Registry is specific to the GridSpace Virtual Laboratory, the concept of utilizing a registry for storing credentials and data source parameters is general and can be applied to other complex data systems. Practical implications of our technique make it an attractive option for implementing data access libraries.

Technology

The GridSpace Engine (GSEngine) -- the execution environment of the GridSpace VL, apart from many other subsystems, incorporates the Data Access Client (DAC2) which is responsible for data access. One of DAC2 components – DSR Connectivity -- equips DAC2 with means of accessing DSR data. Another component -- DACConnector – exposes a public user API accessible from GridSpace scripts and acts as the main entry point for the DAC2 subsystem. The aforementioned public API was expanded to facilitate access to gLite storage and LFC catalogues.

In addition to changes in existing GSEngine modules, two components have been added to GSEngine: the LFC DS client library which is a dedicated Java client capable of contacting the LFC DS Server (an external service acting as a gateway to the EGEE Grid infrastructure) and the LFC DS connector – a JRuby module providing Grid data access and file management capabilities. The LFC DS client and LFC DS Server communicate using two dedicated tunnels, for command and data streaming respectively. As most gLite UI machines are protected by firewalls, there is no easy way of exposing a service running on a gLite UI; thus, tunneling is necessary. The only exception is when both GSEngine and LFC DS Server run on the same machine with access to EGEE. The final component -- LFC DS Server -- is a service that stores all dependencies required to access EGEE storage. By offloading it to a standalone server and not incorporating it into GSEngine, platform independence has been preserved. Moreover, since the client library does not have any GSEngine dependencies, the LFC DS Server can be accessed from any Java program, not only from GSEngine.

Conclusions

The new set of methods available to users is deliberately uncomplicated; however, if any new features are needed (e.g. the ability to set access rights in LFC directories), the LFC DS Server design yields itself to extensions. Moreover, both the LFC DS Server and the Java client library are so generic that they do not have to be used with GSEngine -- they can be used separately in a

different Java project that would benefit from the presented functionality.

To sum up, we have prepared solutions, which will not have to be duplicated by users of EGEE storage services and the virtual laboratories built on top of EGEE resources. Our API is straightforward and while GridSpace requires basic familiarity with scripting languages, easy access to Grid storage resources means that computational scientists are free to focus on domain problems.

Future work carried out on the project includes supplying finer-grained security as well as extending the system to cover additional types of data sources.

References

[1] M. Bubak, M. Malawski, T. Gubała, M. Kasztelnik, P. Nowakowski, D. Harężlak, T. Bartyński, J. Kocot, E. Ciepiela, W. Funika, D. Król, B. Baliś, M. Assel, A. Tirado-Ramos, Virtual Laboratory for Collaborative Applications, in: M. Cannataro (Ed.), Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine and Healthcare, IGI Global, 2009.
[2] P. Sloot, P. Coveney, G. Ertaylan, V. Müller, C. Boucher, M. Bubak, HIV decision support: from molecule to man, Philosophical Transactions A 367 (1898) (2009) 2691.
[3] The GREDIA Consortium, The GREDIA Project Grid Enabled Access to Rich Media Content, in: M. Bubak, M. Turała, K. Wiatr (Eds.), Proceedings of Cracow Grid Workshop - CGW'07, October 2007, ACC Cyfronet AGH, Krakow, Poland, 2007.
[4] M. Sterzel, T. Szepieniec, D. Harężlak, Grid Web Portal for Chemists, in: EGEE User Forum, Catania, Italy, 2009, presentation slides.
[5] J. Kitowski, Structure and Status of National Grid Initiative in Poland, in: M. Bubak, M. Turała, K. Wiatr (Eds.), Proceedings of Cracow Grid Workshop - CGW'08, October 2008, ACC-Cyfronet AGH, Krakow, Poland, 2008.
[6] L. Abadie, P. Badino, J.-P. Baud, J. Casey, A. Frohner, G. Grosdidier, S. Lemaitre, G. Mccance, R. Mollon, K. Nienartowicz, D. Smith, P. Tedesco, Grid-Enabled Standards-based Data Management, in: Mass Storage Systems and Technologies, 2007. MSST 2007. 24th IEEE Conference on, 2007, pp. 60–71. doi:10.1109/MSST.2007.4367964.
[7] D. A. Venton, The LHC Computing Grid in the starting blocks, CERN Bulletin (Issue No. 02/2010).