OGSA-DQP: A Service for Distributed Querying on the Grid

M. Nedim Alpdemir[1], Arijit Mukherjee[2], Anastasios Gounaris[1], Norman W. Paton[1], Paul Watson[2], Alvaro A.A. Fernandes[1], and Desmond J. Fitzgerald[1]

[1] Department of Computer Science
University of Manchester
Oxford Road, Manchester M13 9PL
United Kingdom

[2] School of Computing Science
University of Newcastle upon Tyne
Newcastle upon Tyne NE1 7RU
United Kingdom

Abstract. OGSA-DQP is a distributed query processor exposed to users as an Open Grid Services Architecture (OGSA)-compliant Grid service. This service supports the compilation and evaluation of queries that combine data obtained from multiple services on the Grid, including Grid Database Services (GDSs) and computational web services. Not only does OGSA-DQP support integrated access to multiple Grid services, it is itself implemented as a collection of interacting Grid services. OGSA-DQP illustrates how Grid service orchestrations can be used to perform complex, data-intensive parallel computations. The OGSA-DQP prototype is downloadable from www.ogsadai.org.uk/dqp/. This demonstration aims to illustrate the capabilities of OGSA-DQP prototype via a GUI Client over a collection of bioinformatics databases and analysis tools.

1 Distributed Query Processing on the Grid

Both commercial and scientific applications increasingly require access to distributed resources. Grid technologies have been introduced to facilitate efficient sharing of resources in a heterogeneous distributed environment. Service-oriented architectures are perceived to offer a convenient paradigm for resource sharing through resource virtualisation, and the Open Grid Services Architecture/Infrastructure (OGSA/OGSI) [4] is emerging as the standard approach to providing a service-oriented view of Grid computing. Taken together, these developments have highlighted the need for middleware that provides developers of user-level functionalities with a more abstract view of Grid technologies.

From its inception, Grid computing has provided mechanisms for data access that lie at a much lower level than those provided by commercial database technology. However, data in the Grid is likely to be at least as complex as that found in current commercial environments. Thus, high-level data access and integration services are needed if applications that have large amounts of data with complex structure and complex semantics are to benefit from the Grid.

2 OGSA-DQP

OGSA-DQP [1] is essentially a high-throughput distributed data-flow engine that relies on a service-oriented abstraction of grid resources and assumes that data sources are accessible through service-based interfaces. OGSA-DQP delivers a framework that

- supports declarative queries over *Grid Database Services* (GDSs) and over other web services available on the Grid, thereby combining data access with analysis;
- adapts techniques from parallel databases to provide implicit parallelism for complex data-intensive requests;
- automates complex, onerous, expert configuration and resource utilisation decisions on behalf of users via query optimisation;
- uses the emerging standard for GDSs to provide consistent access to database metadata and to interact with databases on the Grid; and
- uses the facilities of the OGSA to dynamically obtain the resources necessary for efficient evaluation of a distributed query.

OGSA-DQP uses the reference implementation of OGSA/OGSI, viz., Globus Toolkit 3 (GT3) [3], which implements a service-based architecture over virtualised resources referred to as Grid Services (GSs). OGSA-DQP also builds upon the reference implementation of the GGF Database Access and Integration Services (DAIS) standard, viz., OGSA-DAI [2].

OGSA-DQP provides two services to fulfil its functions: The *Grid Distributed Query Service (GDQS)* and the *Grid Query Evaluation Service (GQES)*. The GDQS provides the primary interaction interfaces for the user, collects the necessary metadata and acts as a coordinator between the underlying query compiler/optimiser engine and the GQES instances. GQES instances are created and scheduled dynamically, to evaluate the partitions of a query constructed by the optimiser of the GDQS.

3 The Demonstration

The demonstration illustrates, using a GUI Client, how OGSA-DQP can be employed to orchestrate distributed services for achieving data retrieval and data analysis in a single framework via a declarative query language. The demonstration set-up consists of three distributed database servers, two of which are located in Manchester, UK, and one of which is located in Newcastle, UK. In addition, a analysis web service is deployed on a server in Manchester. It is, however, possible to extend the set of resources with other available data sources and web services if required. The client is implemented in Java and accesses the GDQS via SOAP.

The demonstration shows that a typical query session using OGSA-DQP starts with a set-up phase where the client submits a list of resources (i.e.

data resources and web services) that are anticipated to be required for subsequent OQL queries. The demonstration then illustrates that a query session can continue with multiple query requests to the same GDQS instance. As illustrated in Figure 1 (a), for each request a GDQS instance compiles, optimises, partitions and schedules the query to generate a distributed query plan optimised for specific requirements of the submitted query. It then commands the creation of GQESs as stipulated by the partitioning and scheduling decided on by the compiler (Figure 1 (b), interaction 1), and coordinates the GQESs into executing the plan, which effectively constructs a tree-like data flow system with the GDQS instance at the root, GDS instances at the leaf and a collection of GQESs in the middle (Figure 1 (b), interactions 2-3).

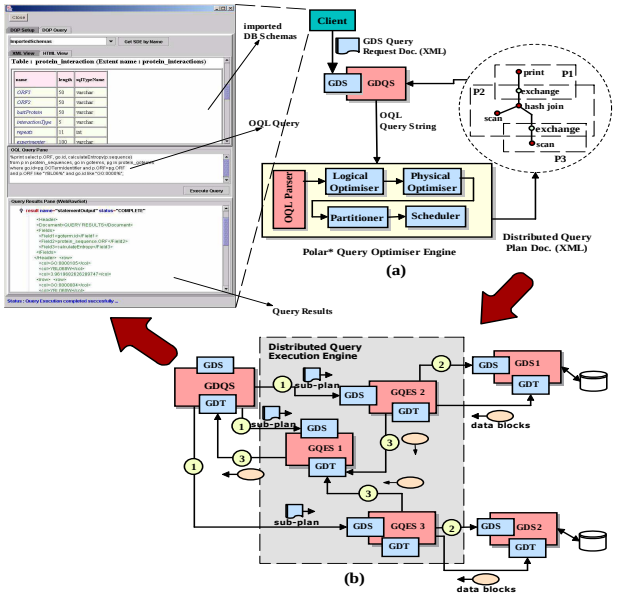

Fig. 1. Execution of a Query

Acknowledgements — The work reported in this paper has been supported by the UK e-Science Programme through the OGSA-DAI and DAIT projects, the myGrid project, and by the UK EPSRC grant GR/R51797/01. We are grateful for that support.

References

1. M. Alpdemir, A. Mukherjee, N. W. Paton, P. Watson, A. A. Fernandes, A. Gounaris, and J. Smith. Service-based distributed querying on the grid. In *the Proceedings of the First International Conference on Service Oriented Computing*, pages 467–482. Springer, 15–18 December 2003.
2. A. Anjomshoaa et al. The design and implementation of grid database services in ogsa-dai. In S. J. Cox, editor, *Proceedings of UK e-Science All Hands Meeting Nottingham*. EPSRC, 2–4 September 2003.
3. T. Sandholm and J. Gawor. Globus Toolkit 3 Core A Grid Service Container Framework. Technical report, 2003. www-unix.globus.org/toolkit/3.0/ogsa/docs/.
4. S. Tuecke et al. Open Grid Service Infrastructure (OGSI). Technical report, OGSI-WG, Global Grid Forum, 2003. Version 1.0.