

Service-based Distributed Query Processing on the Grid

“Imagination is the beginning of creation. You imagine what you desire, you will what you imagine and at last you create what you will.” – George Bernard Shaw

The vision of a Grid where computers, storage and other computational resources are connected to enable mutual sharing, is today becoming a reality. These computing resources can be envisioned as services, which can be created and destroyed on demand. This has led us to the Open Grid Services Architecture, which is merging with the world of web services.

An important service requirement common to many applications is the ability to access and integrate disparate data resources of different types at different sites. This is now possible through the OGSA-DAI (Open Grid Services Architecture – Data Access and Integration) services and the OGSA-DQP (Open Grid Services Architecture – Distributed Query Processor).

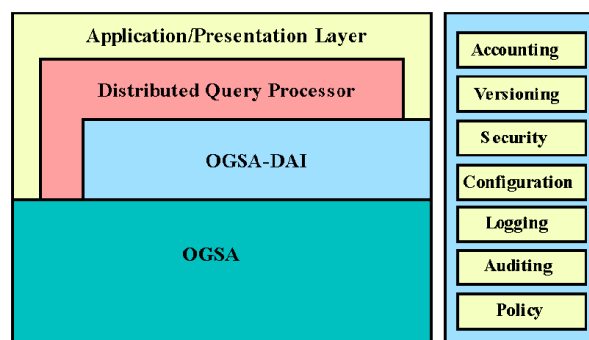
As an example, let us consider a use case from the ^{my}Grid e-Science Pilot Project, which uses a preliminary version of the Distributed Query Processor. A ^{my}Grid bioinformatician may typically want to search for information using very complex conditions, such as:

Find information about all the genes in the GO database whose accession number matches with those in the AffyMapper database, and for which the Probe Set IDs in the AffyMapper database matches with the Probe Set IDs in the MicroArray database where the entries satisfy the “Graves Disease” condition.

This query effectively spans over three different databases, possibly at three

different sites, on three different hosts. Prior to OGSA-DQP, bioinformaticians would require detailed knowledge of the databases being queried and a great deal of manual and error-prone data processing would be needed to integrate the data. In the Grid Services view, the user would submit the query to a service, which behind the scenes would access data from disparate data sources and operate on them in parallel on several machines to maximize efficiency, then return the result. The OGSA-DQP service provides this functionality.

The North-East Regional e-Science Centre has been working closely with e-Sci-



ence North West on service-based distributed query processing for the Grid. The resulting distributed query processor, OGSA-DQP:

- Supports queries over databases wrapped by *OGSA-DAI Grid Data Services* and over other services available on the Grid, thereby combining data access with analysis
- Uses the OGSA facilities to dynamically obtain the resources necessary for efficient evaluation of a distributed query
- Adapts techniques from parallel databases to provide implicit parallelism for complex data-intensive re-

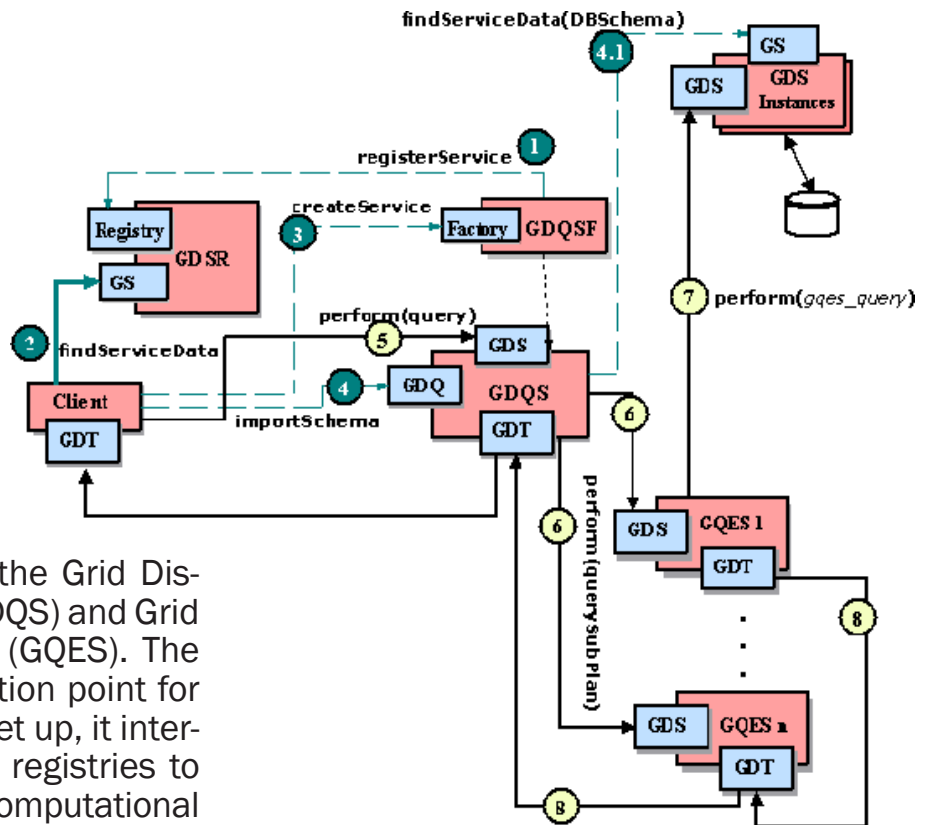
quests so that parts of the query can be executed in parallel on different hosts

- Uses the emerging standard for GDSs to provide consistent access to database metadata and to interact with databases on the Grid

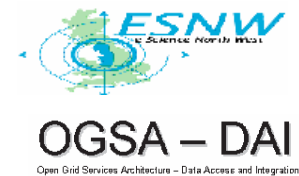
OGSA-DQP extends the OGSA-DAI architecture with two new services – the Grid Distributed Query Service (GDQS) and Grid Query Evaluation Service (GQES). The GDQS is the main interaction point for clients. When a GDQS is set up, it interacts with the appropriate registries to obtain the metadata and computational resource information it needs to compile, optimise, partition and schedule distributed query execution plans over multiple execution nodes on the Grid. The implementation of the GDQS builds on previous work from the Polar* distributed query processor for the Grid by encapsulating its compilation and optimisation functionality.

GQES instances are created by the GDQS based on query plans generated by the query compiler, optimiser and scheduler. This service is otherwise invisible to the external world. Each GQES instance evaluates a partition of the query execution plan assigned to it by a GDQS. The set of GQES instances form a tree through which the data flows from leaf GQESs which interact with GDSs, up the tree to reach its destination.

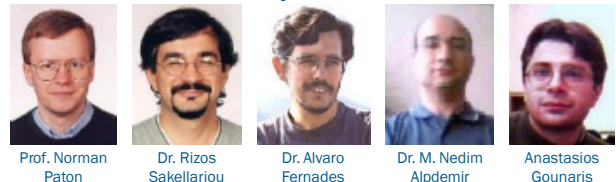
The following diagram depicts the execution flow for the OGSA-DQP system. The dotted lines denote the interaction during the initialisation phase and the solid lines denote the interactions during the execution phase:



The distributed query processor has been designed and implemented as a collection of inter-related Grid services. The OGSA-DQP system is available from: <http://www.ogsa-dai.org.uk>.



Manchester University



University of Newcastle

