Contribution ID: **51** Type: **Oral contribution**

GDSE: A new data source oriented computing element for Grid

Thursday, 2 March 2006 14:00 (20 minutes)

1. The technique addressed in connection with concrete use cases
   In a GRID environment the main components that manages the jobs life are the Grid Resource Framework
   Layer, the Grid Information System Framework and the Grid Information Data Model. Since the job life is
   strongly coupled with its computational environment then the Grid middleware must be aware of the specific
   computing resources managing the job. Until now, only two types of computational resources, the hardware
   machines and some batch queueing systems, have been taken into account as a valid Resource Framework
   Layer instances. However different types of virtual computing machines exist such as the Java Virtual Machine,
   the Parallel Virtual Machine and the Data Source Engine (DSE). Moreover the Grid Information System and Data
   Model have been used for representing hardware computing machines, never considering that a software
   computational machine is even a resource that can be well represented. This work addresses the
   extension of the Grid Resource Framework Layer, of the Information System and of the Data Model so that a
   software virtual machine as a Data Source Engine is a valid instance for a Grid computing model, namely the
   so called Grid-Data Source Engine (G-DSE). Once the G-DSE has been defined, a new Grid element, namely the
   Query Element (QE) can be in turn defined; it enables the access to a Data Source Engine and Data Source,
   totally integrated with the Grid Monitoring and Discovery System and with the Resource Broker.
   The G-DSE has been designed and set up in the framework of the GRID.IT project, a multidisciplinary Italian
   project funded by the Ministry of Education, University and Research; the Italian astrophysical community
   participates to this project by porting on Grid three applications, one of them addressed to the extraction of
   data from astrophysical databases and their reduction by exploiting resources and services shared on the
   available INFN Grid infrastructure whose middleware is LCG based. The use case we envisaged and sketched
   out for this application reflects the typical way astronomers work with. Astronomers typically require to 1)
   discover astronomical data that reside on astronomical databases spread worldwide; this discovery process is
   driven through a set of metadata fully describing the data the user looks for; 2) if data are found in some

archive on the network they are retrieved and processed through a suite of appropriate reduction software
tools; data can also be cross-correlated with similar data residing elsewhere or just acquired by the astronomer; 3) if data the user looks for are not found, the astronomer can decide to acquire them through a
set of astronomical instrumentation or generate them on the fly through proper simulation software tools; 4)
at the end of the data processing phase the user typically saves the results in some database reachable on the
network.
In the framework of our participation to GRID.IT project we realized that the LCG Grid infrastructure based on
Globus 2.4 is strongly computing centric and does not offer any mechanism to access databases in a transparent way for final users. For this reason, after having evaluated a number of possible solutions like
Spitfire and OGSA-DAI, it was decided to undertake a development phase on the Grid middleware to make it
able to fully satisfy our application demands. It is worth to note here that a use case like that described above
is not peculiar of the astrophysical community only, rather it is applicable to other disciplines where access to
data stored in complex structures like database represent a factor of key importance.
Within the GRID.IT project the extended LCG Grid middleware has been extensively tested proving that the
solution under development makes the Grid technology able to fully meet the requirements of typical astrophysical application.
The G-DSE is currently in a prototypal state; further work is needed to refine it and bring it in a production
state. Once the Grid middleware has been enhanced through the inclusion of the G-DSE, the new QE can be
set up. The QE is a specialized CE able to interact, making use of G-DSE capabilities, with databases looking
them as embedded resources within the Grid, like a computing resource or a disk resident file. The QE is able
to process and handle complex workflows that foresee both the usage of traditional Grid resources as well as
the new ones; database resources in particular may be seen and used as data repository structures and even
as virtual computing machines to process data stored within them.

2. Best practices and application level tools to exploit the technique on EGEE
   A suite of tools are currently in the process of being designed and set up to make easy for applications to use
   the functionalities and capabilities of a G-DSE enabled Grid infrastructure. Such tools are mainly thought to
   help users in preparing the JDL scripts able to exploit the G-DSE capabilities and, ultimately, the functionalities offered by the new Grid QE. The final goal however is to offer to final users graphical tools to
   design and sketch out their workflows to be passed on to the QE for their analysis and processing. A precondition, obviously, to achieve these results is to have the G-DSE, and then the QE fully integrated in the
   Grid middleware used by EGEE.
3. Key improvements needed to better exploit this technique on EGEE
   The current prototype of the G-DSE is not included yet in the Grid middleware flavours the EGEE infrastructure
   is based on. The test phase carried out on the G-DSE prototype so far has made use of a parallel test bed Grid
   infrastructure set up thanks to the collaboration between INFN and INAF. Such parallel infrastructure is made
   of a BDII and of a RB on which the modified Grid components constituting the G-DSE have been mounted. The
   mandatory precondition to make use of the G-DSE, therefore is its inclusion (i.e. the modified components of
   the Grid middleware) in the Grid infrastructure used by EGEE.
4. Industrial relevance

The G-DSE has been originally thought to solve a specific problem of a scientific community and the analysis
of new application fields has been focussed so far in the scientific research area.
Because G-DSE however represents a general solution to make of any database an embedded resource of the
Grid, quite apart from the nature and kind of data contained within it, it is natural for the G-DSE to extend its
applicability even in the field of industrial applications whenever the access to complex data structures is a
crucial aspect.

Primary authors: Mr AMBROSI, Edgardo (INFN - CNAF); Dr TAFFONI, Giuliano (INAF - SI)

Co-authors: Mr BARISANI, Andrea (INAF - SI); Prof. GHISELLI, Antonia (INFN - CNAF); Prof. PASIAN, Fabio (INAF - SI); Dr CLAUDIO, Vuerli (INAF - SI)

Presenter: Dr TAFFONI, Giuliano (INAF - SI)

Session Classification: 2b: Data access on the grid

Track Classification: Data access on the grid