DataGrid

Research and Technological Development for an International Data GRID

Abstract

The aim of the European DataGrid (EDG) project is to open up a new world of scientific exploration, by providing software solutions for distributed computation and analysis of large-scale databases. Next generation of scientific exploration requires intensive computation and analysis of shared large-scale datasets across widely distributed scientific communities. EDG addresses this problem by building on emerging Grid technologies, developing the components essential for the implementation of a large-scale data and computational Grid. Very encouraging results have already been achieved in terms of the major goals of the project, which are the demonstration of the practical use of computational and data Grids by the high energy physics, bio-informatics and earth observation communities. A production quality testbed has been set up at a number of EDG sites, which provides a set of common shared services and tools available to all authorized users. A number of Virtual Organizations have been defined for the various research groups involved in the project, and very useful feedback has been collected when scientific applications have been run to measure the performance of the testbed while producing real science. Although a lot of work remains to be done, the validity of the Grid concept and operation has been demonstrated, providing a solid base for the future Grid infrastructure deployment program in Europe.

Objectives

The objective of the DataGrid project is to support advanced scientific research within a Grid environment, offering capabilities for intensive computation and analysis of shared large-scale datasets, from hundreds of terabytes to petabytes, across widely distributed scientific communities. Such requirements are emerging in many scientific disciplines, including physics, biology, and earth sciences.

Testbed

This first testbed deployment was achieved towards the end of 2001 when the first release of the EDG software was deployed and successfully validated. The project has been congratulated "for exceeding expectations" by the reviewers on March 1st 2002, during the first official EU review. The second testbed deployment was successfully performed at the beginning of 2003. At the time of the 2003 European Union review there were 12 sites participating in the production testbed

(see figure), and since then the number has roughly doubled including one site in the United States and one in Taiwan.
The testbed provides significant computing and storage resources to a community of approximately 500 users from thirteen different virtual organizations. A separate development testbed addresses the need for rapid testing and prototyping of the EDG middleware. The reference site for the EDG collaboration is at CERN, where, before any official version of the middleware is released, the initial testing of the software is performed and the main functionalities are proven before distribution.

Applications

The past year has seen significant achievements in the use of EDG middleware by the user communities involved in the project. To make well focused evaluations of the performance of the various releases of the software, EDG/experiment Task Forces were setup at CERN with high energy physics experiments such as ATALS and CMS. The work of the Task Forces helped the development and reconfiguration of the existing series 1 middleware, prior to moving to EDG2 in August 2003. Representatives of the HEP experiments have also been working together with representatives from bio-informatics and earth observation, the other scientific fields supported by DataGrid. The main goal is to identify, detail and prepare specific applications to test the DataGrid services. For earth observation, the main objective is to provide processing power to allow data mining and systematic processing of long time series of data, such as the atmospheric ozone data coming from the ERS GOME satellite instruments. In the biomedical field, 10 applications have been developed which cover a wide range of possible use for the EDG middleware, from the analysis of 3D structure of proteins in biology to mammograms analysis in medicine.

Success Stories

The European Data Grid project has already achieved many of the goals stated at the time of the project conception three years ago. A production quality distributed computing environment has been demonstrated by the EDG testbed, which will now be enriched in functionality, further improved in reliability and extended both geographically and in terms of aggregate CPU power and storage capacity. The community of users has already successfully validated the use of a large set of applications, ranging from High Energy Physics to Bio-Informatics and Earth Observation. EDG software will be used on the large scale production facility that is being setup for the analysis of data that will be produced by the new CERN accelerator (LHC). More developments are currently ongoing to extend the range of functionality covered by the middleware. Collaboration has been established via the GRIDSTART initiative with the other ten existing EU funded Grid projects. In particular, the EU CrossGrid project will exploit DataGrid technologies to support a variety of applications, all demanding guaranteed quality of service. Collaboration with similar Grid projects in the US is being pursued in collaboration with the sister project EU DataTAG.

Project name:
DataGrid

Contract no.:
IST-2000-25182

Project type:
RTD

Start date:
01/01/2001

Duration:
39 months

Total budget:
€ 12,822,960

Funding from the EC:
€9,227,506

Total effort in person-months:
3907

Website:
http://eu-datagrid.web.cern.ch/eu-datagrid/

Contact person:
Dr. Fabrizio Gagliardi
email: Fabrizio.Gagliardi@cern.ch
tel.: +41-22-7672374
fax.: +41-22-7677155

Project participants:

CEA	FR
CERN	Int.Org
CESNET	CZ
CNR	IT
CNRS	FR
CS SI	FR
Datamat	IT
ESA-ESRIN	Int.Org
EVG HEI UNI	D
FOM	NL
IBM	UK
IFAE	ES
INFN	IT
ITC-IRST	IT
KNMI	NL
MTA-SZTAKI	HU
NFR	S
PPARC	UK
SARA	NL
UH	SF
VR	S

ZIB D

Keywords:
Information processing
Information systems
Scientific research
Telecommunication

Collaboration with other EC funded projects:
CrossGrid
Damien
DataTAG
Eurogrid
Géant
GridSTART

IST - Research Networking - Research on Networks – Grids