Subject: **Project Presentation (textual section)**

Author: **Mauro Draoli, Gianfranco Mascari, Roberto Puccinelli**

Diffusion: **This presentation will be put on the project portal and used as an official general project description whenever needed**

Information: **General description of the DataGrid project to interested non technical people**
Version 1_1, 12 July 2001
Includes WP10 and WP8 contributions, Ben Segal and Mark Parsons suggestions. The presentation is now structured according to the project structure
Includes comments of Mark Parsons (13 July 2001)

Project Presentation

Overview

The European DataGrid is a project funded by the European Union with the aim of setting up a computational and data-intensive grid of resources for the analysis of data coming from scientific exploration. Next generation science will require co-ordinated resource sharing, collaborative processing and analysis of huge amounts of data produced and stored by many scientific laboratories belonging to several institutions.

The main goal of the DataGrid initiative is to develop and test the technological infrastructure that will enable the implementation of scientific “collaboratories” where researchers and scientists will perform their activities regardless of geographical location. It will also allow interaction with colleagues from sites all over the world as well as the sharing of data and instruments on a scale previously unattempted. The project will devise and develop scalable software solutions and testbeds in order to handle many PetaBytes of distributed data, tens of thousand of computing resources (processors, disks, etc.), and thousands of simultaneous users from multiple research institutions.

The DataGrid initiative is led by CERN, the European Organization for Nuclear Research, together with five other main partners and fifteen associated partners. The project brings together the following European leading research agencies: the European Space Agency (ESA), France's Centre National de la Recherche Scientifique (CNRS), Italy's Istituto Nazionale di Fisica Nucleare (INFN), the Dutch National Institute for Nuclear Physics and High Energy Physics (NIKHEF) and UK's Particle Physics and Astronomy Research Council (PPARC). The fifteen associated partners come from the Czech Republic, Finland, France, Germany, Hungary, Italy, the Netherlands, Spain, Sweden and the United Kingdom.

DataGrid is an ambitious project. Its development benefits from many different kinds of technology and expertise. The project spans three years, from 2001 to 2003, with over 200 scientists and researchers involved. The EU funding amount to 9.8 million euros.

The first and main challenge facing the project is the sharing of huge amounts of distributed data over the network infrastructure which is currently available. The

DataGrid project does not start from scratch in this challenge: it relies upon emerging computational GRID technologies that are expected to make feasible the creation of a giant computational environment out of a distributed collection of files, databases, computers, scientific instruments and devices. The GRID vision is thoroughly described in a recent book by Ian Foster and Carl Kesselman, which also provides a good summary of the state of the art on meta-computing.

The DataGrid project is divided into twelve Work Packages distributed over four Working Groups: Testbed and Infrastructure, Applications, Computational & DataGrid Middleware, Management and Dissemination.

The DataGrid Testbed

The project will place a major emphasis on providing production quality testbeds, using real-world applications with real data drawn primarily from three scientific areas – high-energy physics, biology and Earth observation. These areas offer complementary data models that allow the demonstration of the cross-field potential of the Data Grid.

The effectiveness of the developed technologies will be demonstrated through the large-scale deployment of end-to-end applications actually used by scientists. The experiments will benefit from the availability of a grid of cooperating data processing centres belonging to many research institutions spread across Europe. The necessary broadband research networking infrastructure will be provided by other European Research initiatives.

From a technological point of view, the challenge is to demonstrate the possibility of building very large clusters of distributed resources out of low-cost computing commodities. This approach towards commodity-based solutions has been successfully applied by most of the project partners over the past ten years within their production environments.

The testbed will also ensure that the output of the project will contribute to the basic requirements of performance and reliability to comply with real-world computing standards.

Applications Areas

The three real data intensive computing applications areas covered by the project are:

- High Energy Physics (HEP), led by CERN,
- Biology and Medical Image processing, led by CNRS (France),
- Earth Observations (EO) led by the European Space Agency.

High Energy Physics applications

One of the main challenges for High Energy Physics is to answer longstanding questions about the fundamental particles of matter and the forces acting between them. In particular the goal is to explain why some particles are much heavier than others, and why particles have mass at all. The answer could reside in an all pervading presence called the “Higgs field”, but at the moment we have no evidence of its existence. To that purpose CERN is building the Large Hadron Collider, the most powerful particle accelerator ever conceived, that will provide data related to such interactions at a tremendous output rate. The DataGrid will provide the solution for storing and processing such huge amounts of data. A multi-tiered, hierarchical computing model will be adopted to share data and computing efforts among multiple institutions. The Tier-0 centre will be located at CERN and will be linked by high speed networks to approximately ten major Tier-1 data processing centres. These will fan out the data to a large number of smaller ones (Tier-2).

Biology and Medical applications

The storage and exploitation of genomes and the huge flux of data coming from post-genomics puts growing pressure on computing and storage resources within existing physical laboratories.

Medical images are currently distributed over medical image production sites (radiology departments, hospitals). Although there is today no standard for sharing data between sites, there is an increasing need for remote medical data access and processing.

The DataGrid project's biology testbed will provide the right platform for new algorithms on data mining, databases, code management, graphical interface tools and will facilitate sharing of genomic and medical imaging databases for the benefit of international cooperation and health care.

Earth Observation applications

The European Space Agency missions involve the download, from space to ground, of about 100 Gigabytes of raw images per day. Dedicated ground infrastructures have been set up to handle the data produced by instruments onboard the satellites.

The analysis of atmospheric ozone data has been selected as a specific testbed for the DataGrid. More generally, the project will demonstrate an improved way to access and process large volumes of data stored in distributed European-wide archives.

The DataGrid Middleware

DataGrid consists of physical resources (computers, disks and networks) and "middleware" software that ensures the access and the co-ordinated use of such resources. Such a software, the "glue" of the DataGrid, is to be developed in collaboration with some of the leading centres of competence in GRID technology, leveraging practice and experience from previous and current GRID initiatives in Europe and elsewhere. The glue role of DataGrid middleware is further highlighted by considering its potential for cross-application within the three different reference areas mentioned above: high energy physics, biology, Earth observation.

The Middleware component of the project coordinates the development of the necessary software modules leveraging, where possible, existing and long tested open standard solutions. Five parallel work packages are defining and implementing the core services of the DataGrid middleware: job scheduling, data management, grid monitoring, fabric management and mass storage management.

The DataGrid project participants are directly collaborating with key members of the Globus project, the main initiative for the development of Grid middleware in the US, and with the GriPhyN project that has recently been funded by the US National Science Foundation. The software produced will extend the state of the art in international, large-scale, data-intensive grid computing into the framework of the Global Grid Forum, the initiative born to group and coordinate US, European and Asian projects.

Each release of the software components will be integrated and made available for installation at the different testbed sites. All tools and middleware developed by the project will be released as Open Software. The open approach is important for two reasons:

- at this early stage of GRID technology development, it is important to be able to discuss freely within the international scientific community;
- considering that the project is limited to basic technology advance and demonstration of the potential of underpinning ideas, the open approach ensures that the project results will be freely available for commercial exploitation by any organisation in the future.

Outreach

A consequence of the project will be the emergence of new modes of scientific exploration, as access to fundamental scientific data is no longer restrained to their producer. While the project focuses on scientific applications, issues of data sharing are germane to many applications with a potential impact on future industrial and commercial activities.

Grid technologies in general will have important social outreaches. The long term goal of this new technology is to provide applications, data, computing and storage power to Grid users on demand.

Possible outreaches are:

- providing access to the most powerful computing facilities to remote and less developed areas;
- increasing the competitiveness of small and medium enterprises, enabling them to access the same advanced computational tools and resources available in large enterprises;
- enhancing the economic development potential and the subsequent life quality in small towns by slowing the devastating tendency to move all industries and highly skill workforces to large industrial sites.

The DataGrid project is contributing to the revolution that is already underway. Involving European scientists and computer experts at the heart of the world-wide GRID initiative will keep Europe at the forefront of the new Information Society development.

This project will also address the ethical aspect of promoting high level education and integration of many different countries from EU and Central as well Eastern European areas. As part of its dissemination activity, a network of proactive industries and research institutes is grouped in a so-called *Industry and Research Forum.* In particular this brings together researchers in different disciplines (Earth Observation, Biology, Physics, and Computer Science) from many EU member states, European Industry and countries such as USA, Japan, Russia, Hungary, Poland, Czech Republic and Israel.