

**Initial development and exploitation of a DataGrid  
infrastructure in Spain in support of the Large Hadron  
Collider and other experiments**

*DRAFT 06/08/2001 17:27*

*Prof. Manuel Delfino*

*Departament de Física, Universitat Autònoma de Barcelona*

*(REAL AUTHOR LIST TO BE INSERTED)*

ABSTRACT OR EXECUTIVE SUMMARY

To be done

## TABLE OF CONTENTS

1	Introduction.....	1
2	Requirements for data-intensive computation of the Large Hadron Collider experiments .....	3
3	Requirements for data-intensive computation of other experiments with Spanish participation.....	3
4	DataGrid infrastructure from the user point of view.....	3
5	DataGrid infrastructure from the provider point of view.....	3
6	Initial phase of DataGrid infrastructure in Spain .....	4
7	Capacity planning for the LHC integration and commissioning phase (2002-2004) .....	5
8	Coordination of Spanish participation in the LHC Computing Grid Project...	5
9	Management, coordination and dissemination .....	5
10	Summary of required resources for the period 2002-2004 .....	5
11	Appendix A: Specific Objectives and Funding Request (Barcelona IFAE)	5
12	Appendix B: Specific Objectives and Funding Request (Madrid CIEMAT)	5
13	Appendix C: Specific Objectives and Funding Request (Madrid UAM) ....	5
14	Appendix C: Specific Objectives and Funding Request (Santander IFCA)	5
15	Appendix D: Specific Objectives and Funding Request (Santiago de Compostela) .....	5
16	Appendix A: Specific Objectives and Funding Request (Valencia IFIC) ...	6
17	Appendix A: Specific Objectives and Funding Request (next institute) .....	6



# **Initial development and exploitation of a DataGrid infrastructure in Spain in support of the Large Hadron Collider and other experiments**

## **1 Introduction**

A number of important developments have taken place in the last few years in the field of scientific computing, stimulated by advances in technology and changes in scientific methodologies. Data-intensive quasi-parallel computing has joined vector super-computing and massively parallel computing in pushing the envelope of high performance computing. Data-intensive computing is expected to explode in the next few years, driven on the one hand by advances in storage and networking technologies and on the other hand by the availability of enormous amounts of scientific data in digital form. A leading adopter of multi-Petabyte scale data-intensive computing will be the application of data analysis of the experiments of the CERN Large Hadron Collider (LHC), due to start operations in 2006. Other applications with lower overall data volumes, but rather complex pattern recognition algorithms, are found in the fields of Bio-Informatics and exploitation of Earth observation data.

Another important development is the so-called Peer-to-Peer computing techniques, which when combined with advances in optical networking such as Dense Wavelength Division Multiplexing and an appropriate security and access control architecture will result in the next generation of distributed computing, that is Information Grids. Grids will allow the transparent support of sophisticated collaborative data processing environments for geographically distributed virtual communities. Although first conceived as a way to create meta-supercomputers, the idea of Information Grids has sparked the imagination of scientists and, perhaps a bit as a surprise, of the general public. The potential of using, even primitively, indexing and directory technologies and P2P protocols over the

public Internet infrastructure to transparently support users in arbitrary locations, has been dramatically illustrated by the (quite controversial) Napster phenomenon.

Perfecting the techniques described above to include appropriate security and access control, and above all developing and deploying the Grid infrastructure to enable delivery of very high throughput in data-intensive environments is the aim of a number of projects, in particular the EU DataGrid and the LHC Computing Grid projects, stemming from the High Energy Physics community. Support for this community, and in particular for the fulfilment of the data analysis needs of the experiments of the LHC, represents an excellent laboratory for data-intensive Information Grid development (or DataGrid development for short), driven by the very real users needs of a worldwide community of motivated scientists with a common goal: The search for the very reason of the existence of mass, the Higgs boson, by the collaborative work of thousands of professors and doctoral students sifting through their experiment's data for signals as rare as 1 in  $10^{12}$ .

Participation by the Spanish HEP community in Grid developments started in late 2000 through involvement in the EU funding request for the DataGrid project, and was strengthened in 2001 by a successful bid on the EU CrossGrid cross-action and the granting of single year Special Action projects by the Ministry of Science and Technology. The experience during 2001 indicates that there is very good potential for high quality contributions from Spain to Grid development in Europe, in the context of solving the enormous challenge of data analysis for the LHC. In order to fully realize this potential, this proposal outlines a three-year project (2002-2004) for the initial development and exploitation of DataGrid infrastructure in Spain, driven by the requirements of the integration and commissioning phase of the LHC experiments, but also capable of supporting data analysis of ongoing physics experiments. Funding of the project will not only benefit the Spanish HEP community, but will also serve to stimulate Grid

developments in Spain in general, with potential benefits for technology transfer to other fields with data-intensive computing requirements.

## **2 Requirements for data-intensive computation of the Large Hadron Collider experiments**

### *2.1 Integration and Commissioning Phase (2002-2004)*

To be filled out based on the LHC Computing Review report and the discussions in Barcelona in January 2002.

### *2.2 Physics Analysis Preparation and Initial Operation Phase (2005-2007)*

To be filled out based on the LHC Computing Review report.

(This section should be kept short but is placed here to be able to fire a “warning shot” to funding authorities about Phase 2 of LHC Computing. It can also mention participation in the Technical Design Report for the LHC Computing Grid Project.)

## **3 Requirements for data-intensive computation of other experiments with Spanish participation**

Possibility to put here CDF, LEP, AMS, Magic, etc.

## **4 DataGrid infrastructure from the user point of view**

Fill in one or two paragraphs explaining the basic concepts of Grid-enabled applications and Grid infrastructure creating the illusion for the user of a very powerful metacomputer. Add some HEP specific details due to the large volume, unpredictable subsets of data an individual user needs.

## **5 DataGrid infrastructure from the provider point of view**

Fill in one or two paragraphs explaining how DataGrids are built-up. Keep short by using analogy to power plants, transformers, transmission lines, given that later there is a detailed description of the actual infrastructure we want to build up.

## 6 Initial phase of DataGrid infrastructure in Spain

The infrastructure for the Spanish DataGrid Zone will be organized according to the architecture shown in Figure 1.

The architectural building blocks of this infrastructure are described in the next sections, starting from those which are expected to be most specific to each physics experiment and ending with those which are expected to be of a quite general nature; this corresponds to reading Figure 1 from the bottom up. The blocks will be organically bound using a Security, Authentication and Access Control Architecture and a Grid Services Architecture in such a way that the infrastructure behaves as a single transparent system which responds to the user demands, as generated by Grid-enabled versions of data analysis and simulation applications from the LHC and other physics experiments.

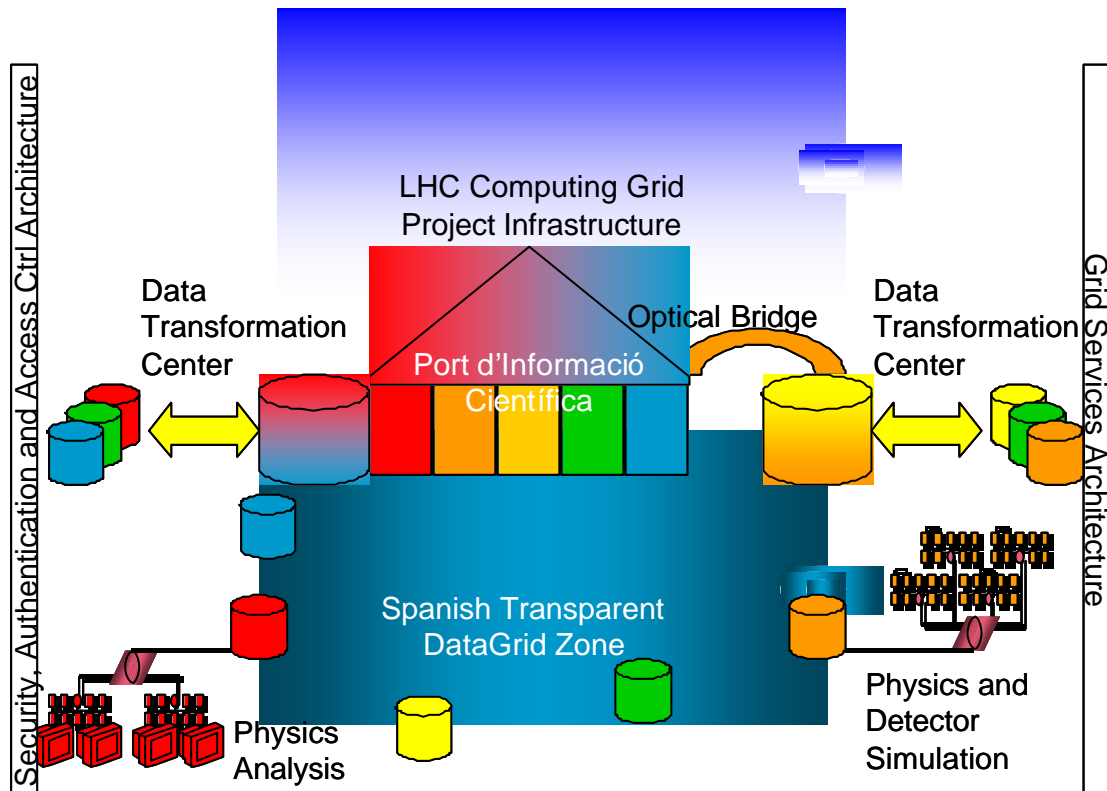


Figure 1: General architecture of the infrastructure of the Spanish DataGrid Zone.

- 6.1 *Software Development Stations*
- 6.2 *Application Configuration and Deployment*
- 6.3 *Data Analysis Stations*
- 6.4 *Detector Simulation Centers*
- 6.5 *Data Transformation Centers*
- 6.6 *The Port d'Informació Científica (PIC)*
- 6.7 *Interface between the Spanish DataGrid Zone and the LHC Petabyte DataGrid Service infrastructure*
- 6.8 *The Gigabit virtual network backbone*
- 7 Objectives, milestones and schedule for the LHC integration and commissioning phase (2002-2004)**
- 8 Relation to EU DataGrid and CrossGrid projects and coordination of Spanish participation in the LHC Computing Grid Project**
- 9 Management, coordination and dissemination**
- 10 Summary of required resources for the period 2002-2004**
- 11 Appendix A: Specific Objectives and Funding Request (Barcelona IFAE)**
- 12 Appendix B: Specific Objectives and Funding Request (Madrid CIEMAT)**
- 13 Appendix C: Specific Objectives and Funding Request (Madrid UAM)**
- 14 Appendix C: Specific Objectives and Funding Request (Santander IFCA)**
- 15 Appendix D: Specific Objectives and Funding Request (Santiago de Compostela)**

**16 Appendix A: Specific Objectives and Funding Request (Valencia IFIC)**

**17 Appendix A: Specific Objectives and Funding Request (next institute)**