EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH
European Laboratory for Particle Physics

Publication
ALICE reference number
ALICE-PUB-2001-48 version 1.0
Institute reference number

Date of last change
2001-10-18

Grid Computing: Concepts and Applications

Authors:

P. Cerello
for the ALICE Collaboration

GRID COMPUTING: CONCEPTS AND APPLICATIONS

Piergiorgio Cerello

I.N.F.N., Sez. di Torino, Via P. Giuria 1, I−10125, Torino, Italy
E−mail: cerello@to.infn.it

ABSTRACT

The challenge of CERN experiments at the Large Hadron Collider (LHC), which will collect data at rates in the range of PBs/year, requires the development of GRID technologies to optimize the exploitation of distributed computing power and the authomatic access to distributed data storage.

Several projects are addressing the problem of setting up the hardware infrastructure of a GRID, as well as the development of the *middleware* required to manage it: a working GRID should look like a set of services, accessible to registered applications, which will help cooperate the different computing and storage resources.

As it happened for the World Wide Web, GRID concepts are in principle important not only for High Energy Physics (HEP): for this reason, GRID developers, while keeping in mind the needs of HEP experiments, are trying to design GRID services in the most general way.

As examples, two applications are described: the CERN/ALICE experiment at the LHC and a recently approved INFN project (GPCALMA) which will set up a GRID prototype between several mammographic centres in Italy.

1. HIGH ENERGY PHYSICS EXPERIMENTS NEED A GRID

The next generation of High Energy Phisics experiments, now in preparation, will operate in the interaction regions of the CERN Large Hadron Collider [1] in a few years from now.
In trying to reach their physics goals, like looking for the Higgs boson and for the Quark Gluon Plasma, these experimental facilities will collect data at an unprecedented rate, up to 3 PB/year each. Most of these data (about 80%) will be raw information coming from the front−end electronics, which must be processed in order to optimize the detector performance and to reconstruct and analyze the events, up to the final physics results.

These numbers exceed by far the expected capability of any single Computing Centre. In addition, the storage capacity and CPU speed improvements are expected to be of the order of a factor of 20 in the next 5−6 years, compared to an increase of about 1,000 times in the storage and computing requirements with respect to the previous generation of HEP experiments. The number of single computing units needed will then increase by about 50 times,and the simple centralized model for the data storage and computing cannot be applied.

Moreover, the LHC Collaborations (ALICE, ATLAS, CMS, LHCb) are very large and involve a total of about 6,000 people, coming from institutions scattered all around the world,

who will need to access reconstructed data for different kinds of analysis. Therefore a distributed model for both the storage and the computing power is mandatory, in order to allow an optimal exploitation and sharing of resources. However, such a model requires the availability of *middleware* tools capable of dealing with typical issues related to a multi−user distributed environment:

- security issues related to the large number of users;
- configuration of a large number of computing farms;
- management of the full data set;
- job submission to local and remote nodes;
- monitoring of the CPU, storage and network loads;
- management of a distributed bookkeeping and calibration database.

The development of GRID *middleware* is already taking place in the framework of various projects (EU−DataGRID [2], GriPhyN [3], PPDG [4]), in Europe or America, started by the HEP community. However, many other GRID potential users are showing interest, among both research institutes and companies. As an example, groups working in Earth Observation and Biology are participating to the EU−DataGRID project. For the HEP community,the timescale to get a full size working GRID system is set by the LHC scheduled starting time in year 2007.

2. DATA AND COMPUTING GRIDs

The GRID approach is based on the assumption that a geographically distributed community, be it a group of researchers or a company based in different cities or countries, needs to facilitate the access to remote data and computing resources in order to improve the effectivity of its activities. In the GRID terminology, this ensemble is called a *Virtual Organization* (VO). However, different virtual organizations are in general interested in different aspects of the problem, having different requirements and goals. The software architects and the *middleware* developers are designing a modular system based on the cooperation of various individual components, called GRID services, as independent as possible of the applications. However, the application software must be interfaced to the GRID Services, and a smart definition of the interface is a key issue for the success of these projects. For this reason, all of them directly involve application experts in the design of the system architecture.

Generally speaking, a GRID can be viewed as a system requiring some actions to define its configuration and able to dynamically process user requests. However, the GRID itself is to be considered as dynamical, though its evolution takes place on a time scale much longer than that set by the execution of user requests. The following subsections are dedicated to a simplified description of the two different kinds of operations: how to set up a GRID, defining its different services, and how the GRID services cooperate to process a user command/job.

2.1 GRID CONFIGURATION

2.1.1 SECURITY

The introduction, for the first time, of the concept of authomatic remote access to distributed data storage and computing power is intrinsically posing a very serious problem of security. Indeed, the mechanism to allow remote users on local systems must meet a high security standard, in particular for communities dealing with private data. The current view is a double level mechanism, based on the identification of the user via a private key issued by a registered Certification Authority (*Authentication*), accepted by all the partners belonging to a given GRID,

followed by an *Authorization* step, managed by the different local sites, which can define the list of accepted certificates and associate them to a list of local user accounts. The services required to manage the GRID users, therefore, consist in a Certification Authority, which could be common to several VOs, and of a User Monitoring Service, which is a part of the overall GRID monitoring system, and will most likely be set up by publishing a tree of user names associated to their Certificate Subjects, with different branches for different VOs. Since GRID resources could be shared by different VOs, each user will be allocated a given amount of resources and an accounting system to monitor their use will be developed.

2.1.2 FARM CONFIGURATION

Beside users, the other relevant component to be properly configured is the set of computing farms belonging to a GRID: in particular, the set of computing elements (CE), on which the operating system, the GRID services and the VO software must be installed and configured. Since the number of CEs can be very large, the VO software is likely to change quite often and new VOs could decide to make use of GRID Services, it is particularly important to develop an authomatic and modular tool, which should also allow an easy update of the farm configuration. Moreover, GRID Services must be developed in such a way to be easily interfaced to existing Resource Management Systems, which define locally the resource sharing between the different CPUs. The same considerations hold for storage elements (SE), whatever the hardware used (disk, tape, CD−rom, DVD) and the Mass Storage Management System being used at a site.

2.1.3 DATA MANAGEMENT

The concept itself of a *Virtual Organization* implies to whole set of data produced by its users must be available and shared by all of them, no matter what is their physical location. Indeed, in case the amount of data cannot be permanently stored at a single centre, it is essential to develop GRID Services that allow their management as a common data set. The present approach is based on the identification of each file with a Logical File Name (LFN), assigned by the GRID, with a part of it defining the VO to which it belongs. The LFN is registered on a so−called Replica Manager Service, and associated to the list of available physical copies, defined by their Physical File Names (PFN). However, the LFN must also be transferred to the user, in order to be stored in the MetaData Catalogue, managed by the VO, which associates the LFN to the set of parameters defining its content.

The distributed data sharing, in order to be effective, requires a minimum number of files to be updated, as in this case it is difficult to keep the consistency between different copies mirrored at different locations.

2.1.4 MONITORING SERVICES

The optimal operation of a GRID is based on the knowledge of the state of its components at all times. Indeed, all the choices concerning data transfers, computing and storage elements selections are based on the availability of the different GRID elements, as well as the network conncetions.

The present view of the Monitoring Service is based on the idea that the different GRID elements will send information about their status to a set of servers which will make it available to all the GRID services and, eventually, to the users themselves.

2.2 GRID OPERATION

The system components described in the previous section must be managed in order to dynamically react to user requests, whether jobs or interactive commands. In other words, a high level Service, known as Resource Broker or Workload Manager, must be available to the registered users in order to allow them to define the input of their jobs and submit them to the GRID.

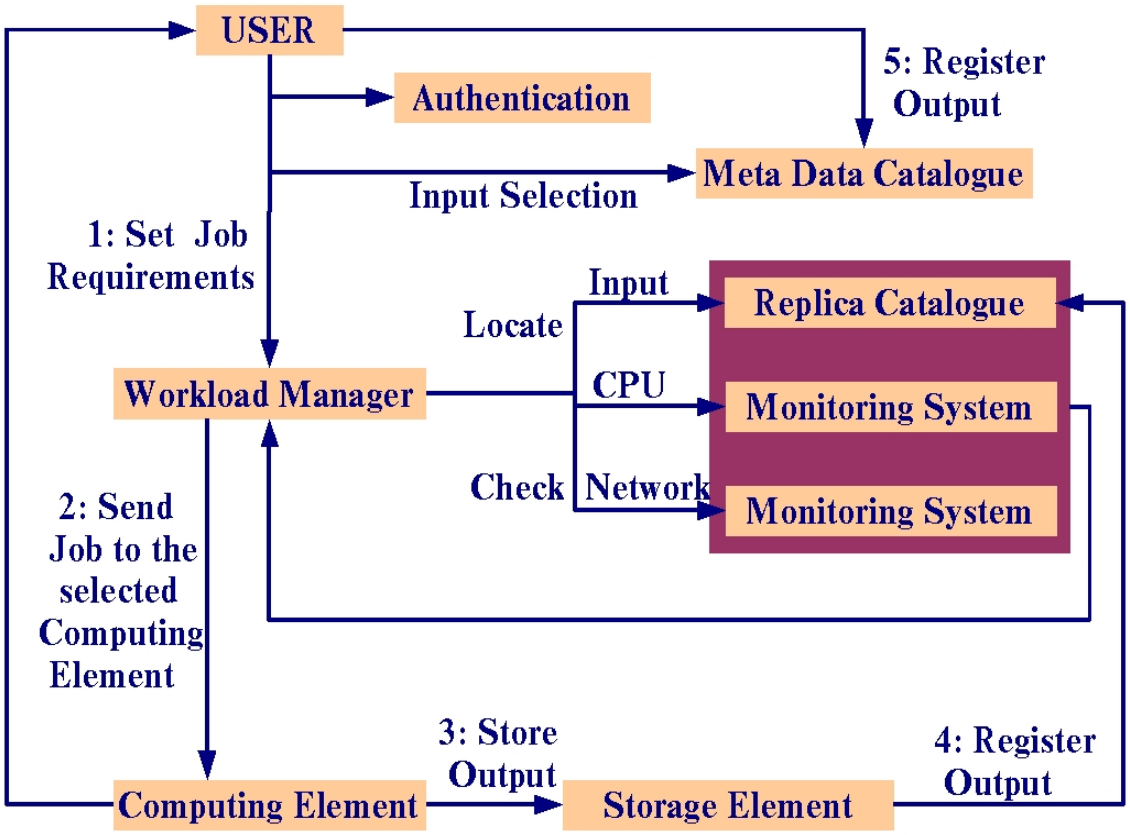

Figure 1: A simplified scheme of the operation flow when a job is submitted to a GRID system.

The simplified operation flow, sketched in Fig. 1, can be summarized as follows:

- the user authenticates himself;
- the user selects the job input, in the form of a list of LFNs, via a query to the MetaData Catalogue;
- the user defines the job input and submits it to the Workload Manager;
- the Workload Manager queries the Replica Catalogue in order to find out the physical locations of the input files, selects the available Computing Elements and checks the network connections, via the Monitoring System;

- the Workload Manager, making use of the collected information, defines which Computing Elements should run the job and eventually moves the input files to their associated Storage Elements; then, the control is transferred to the Computing Elements;
- the CEs execute the job and store its output on the selected Storage Elements, which, in turn, register the new files on the Replica Catalogue;
- the control returns to the user, who registers the output files in the MetaData Catalogue.

3. GRID APPLICATIONS

Several collaborations, mainly in the field of High Energy Physics, are already actively working in the framework of different projects, in Europe, Asia and America. In the following I will focus on two of them, selected according to my personal involvement: the CERN/ALICE experiment [5,6,7] and the INFN/GPCALMA project.

3.1 THE CERN/ALICE EXPERIMENT

The ALICE experiment [5] at the CERN LHC [1] is focused on the search for the Quark Gluon Plasma in heavy ion collisions. The LHC lead beams, at 2.76 TeV/A each, will provide a maximum collision energy of 1,148 TeV, corresponding to expected charged multiplicities as large as 8,000 particles per rapidity unit for central collisions. The data collected in one year of operation (one month for Pb−Pb, ten for p−p interactions) are foreseen in the range of a few PBs. The data acquisition rate will be limited by the maximum bandwidth of 1.25 GB/s. For the Pb−Pb run, with an average event size of 25 MB, the data acquisition rate will be about 50 Hz. The integrated running time is expected in the range of 10^6 s/year, corresponding to $5x10^7$ events/year, and therefore to a raw data amount of 1.25 PB/year. A similar amount will be collected during the p−p run: indeed, the event size will be smaller, but compensated by the higher luminosity and the longer running time. The contribution from the reconstructed data and the analysis objects will be of the order of 500 TB/year, bringing the total data production rate to about 3 PB/year.

The ALICE community is mainly involved in EU−DataGRID; however, US groups which recently joined are evaluating the possibility to collaborate in the GriPhyN development.

The framework for the development of the interface between the ALICE software (AliRoot [6], based on ROOT [8]) and the GRID *middleware* is set by the need to deliver the results of the Physics Performance Report (PPR) by the end of year 2002: the ALICE detector acceptance, efficiency and resolution for the set of interesting physics signals must be evaluated. In order to complete the PPR, the simulation of about 10^4 full central Pb−Pb events, each one of about 2 GB size, corresponding to a total storage of 20 TB, is required. From the point of view of computing power, the simulation and reconstruction of one central Pb−Pb event is equivalent to about $2.25x10^4$ kSI2000 s.

3.1.1 THE ALICE DISTRIBUTED TEST

In order to deal with these numbers, it is important to start now the setting up of a distributed environment, involving at least some of the sites expected to become important ALICE Regional Computing Centres. The ALICE approach to the GRID infrastructure development is taking place along two different mainstreams:

- set up of a working system making use of presently available tools, with the involvement of the sites with relevant storage and computing power already available;
- development of a GRID related infrastructure within the EU−DataGRID project, involving all the ALICE active sites.

The activity carried on up to now consisted in configuring the hardware and software of involved sites and running a distributed production test, driven by a central manager, which, making use of GLOBUS [9], submitted jobs to all the remote sites, getting back the logging information and storing it into a database. In terms of the PPR, this step corresponds to the production of events; the reconstruction and the analysis starting from a distributed input would require the availability of *middleware* tools under development in the EU−DataGRID project.

3.1.2 HARDWARE RESOURCES

Presently available resources amount to a total of:

- about 50 kSI95, mainly at CC−IN2P3, Lyon;
- about 10 TB of disk space and about 150 TB mass storage, managed by HPSS or CASTOR, depending on the site.

Wherever possible, the resources at each site are seen through one single IP node, having access to the complete site storage capacity and acting as interface to the Wide Area Network.
The network connectivity is presently good only for few sites, namely CERN, Lyon, INFN, GSI, NIKHEF and OSU, and it should certainly be improved.

3.1.3 SOFTWARE CONFIGURATION

The basic software configuration is presently defined by the availability of the following tools:

- ALICE simulation software environment (ROOT and AliRoot), installed via either a precompiled distribution or a script which downloads the code from the corresponding CVS servers and compiles it automatically. The kit is being upgraded to meet the EU−DataGRID request of application software configuration via RPM file. Both versions of the kit are available on the WEB [7].
- GLOBUS, in order to be able to qualify as GRID users and submit jobs to remote sites.
- Local Resource Management Systems (LRMS). Many of them are in use (BQS, LSF, PBS), the choice being essentially left to the freedom of the local site administrators; in fact, LRMS are not seen directly, as they are managed by the GLOBUS GRAM.
- Certification Authority environment. Some of the sites have access to their own National Certification Authorities; the others are presently making use of test certificates released by the GLOBUS team. In the medium term, the CERN Certification Authority will issue certificates for all the ALICE users, as they are registered as CERN users.
- MRTG [10], used to monitor the CPU, storage and network load during the production.
- EnginFrame [11], a WEB interface for the access to authentication and job submission services.

3.1.4 BOOKKEEPING

The job bookkeeping is presently controlled by the site acting as manager. It is based on the information stored in the standard output, which is retrieved to the submitting site and parsed.

The collected information is then stored in a DataBase (either MySQL or Oracle) available on the WEB. Work is going on on this item, in order to allow any site running a job to update the bookkeeping database, in view of the reconstruction and analysis phase, when many users will send jobs randomly.

3.1.5 RESULTS

The layout of the distributed system is showed in Fig. 2. There exist essentially three players: the production manager, the local administrator and the generic user. The production manager selects the worker node among the available options, sets up the job input and transfers it to the Run DataBase Server; then, the control goes to the worker node, which retrieves the input from the DataBase Server and executes the job. During the execution, the standard output and the standard error are periodically updated on the DataBase Server. At the end of the job, the permanent output can be transferred to the Mass Storage Systems at CERN or Lyon.

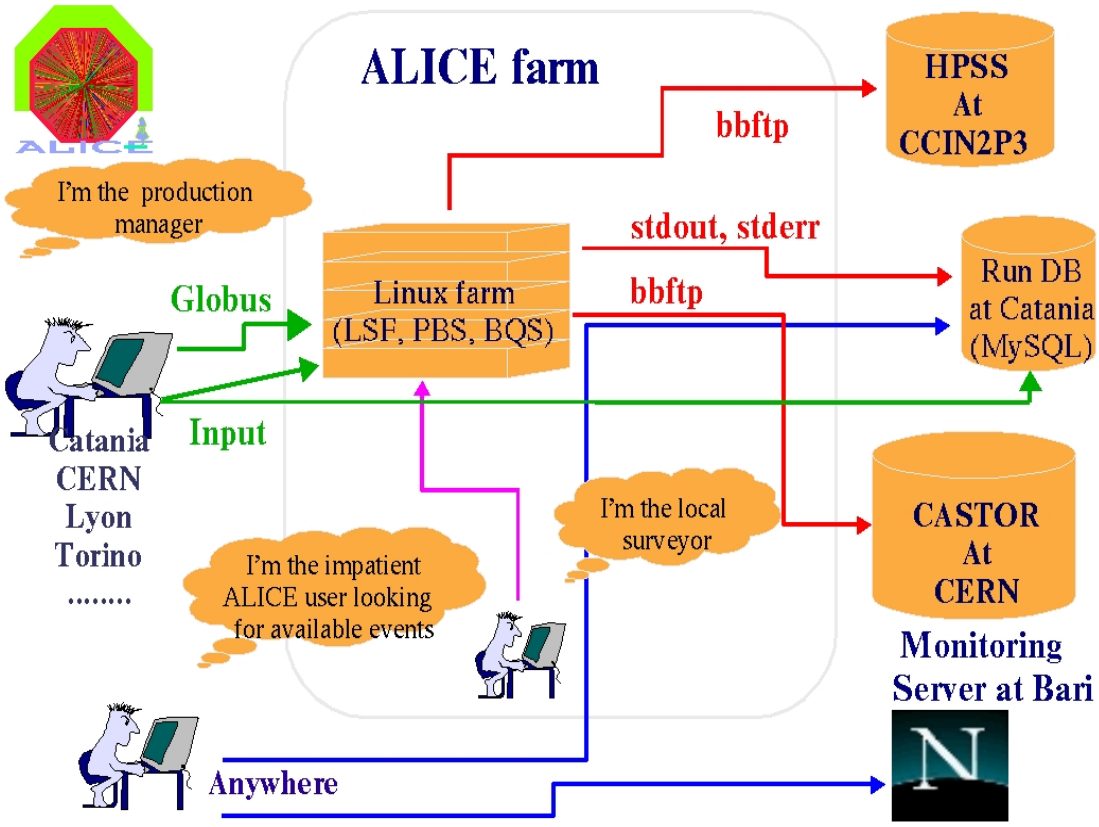

Figure 2: The layout of the ALICE distributed system.

The distributed test was run for a week; about 200 central Pb−Pb events were generated and tracked, producing an output of about 400 GB, which was kept at production sites. However, some data transfer test between Lyon and CERN took place and the interface to HPSS and CASTOR mass storage was tested as well. One of these events was submitted from Torino to Mexico City, on a very slow (266 MHz) machine: it ran for about three days and came

successfully to an end. No crashes were registered during the test, even though the statistics is not significant enough to test AliRoot against rare errors.

The reconstruction (few users) and analysis (many users) of simulated events will require the availability of some GRID Services: a Replica Catalogue for the access to distributed data and a Workload Manager for the load balancing of resources. However, the ALICE timescale fits pretty well with the expected EU−DataGRID releases; therefore, a proper interface of the software to make use of GRID Services could be developed.

3.1.6 LONG TERM PLANNING

The ALICE computing model for the long term is based on the use of GRID Services and the Parallel ROOt Facility (PROOF, part of ROOT). The basic idea, sketched in Fig. 3, is that massive data transfer should be avoided whenever possible; instead, the code to be executed is sent to the working node as a text file. Indeed, it needs not be compiled, as ROOT provides a C++ interpreter. Therefore, an ALICE job is expected to go through the following steps:

- Input data selection, performed via queries to the ALICE Data Catalogue, storing all the basic information related to simulated, raw or reconstructed events. The query ends with a list of Logical File Names (LFNs), which contains the information needed by the GRID to locate the physical files and is given as part of the job description;
- Creation and submission of the job to the GRID Scheduler;
- Analysis script execution, in parallel at all the nodes containing a fraction of the input events;
- Output collection, in the form of a list of LFNs, and update of the ALICE Data Catalogue;
- Local merging of the collected output and final analysis.

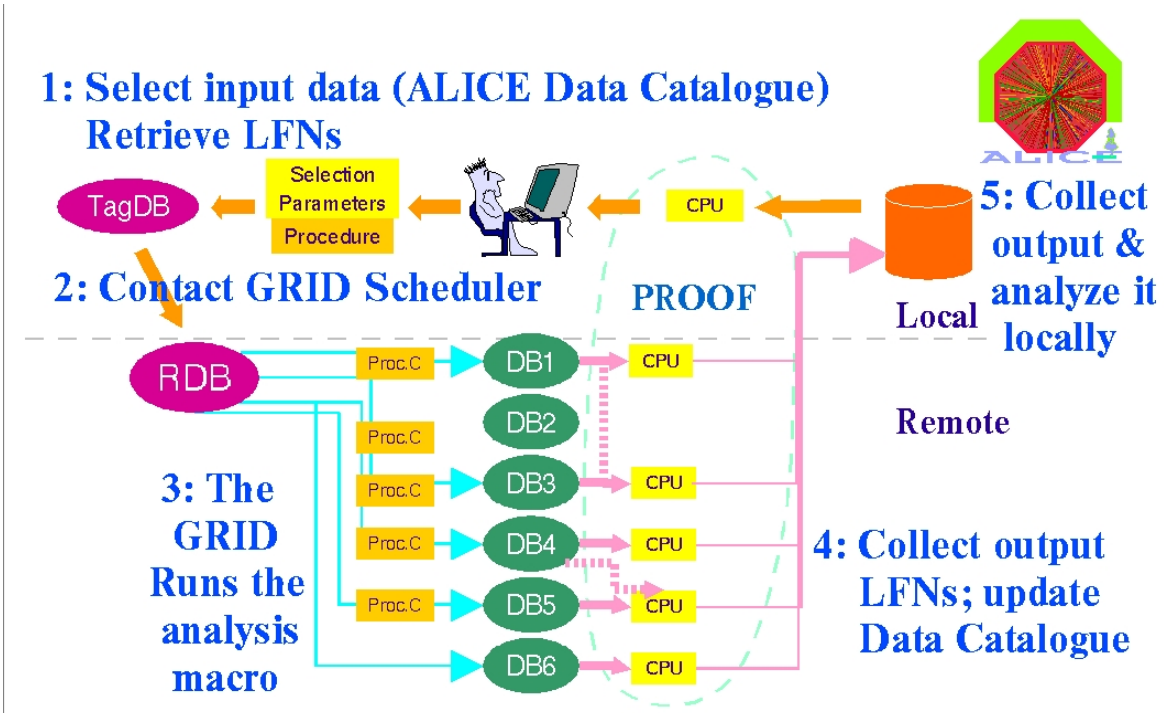

Figure 3: The long term scheme for the ALICE interface to the GRID system.

3.2 THE INFN GPCALMA PROJECT

The INFN CALMA (Computer Assisted Library in Mammography) project set up the largest European database of digital mammographic images (about 5,000), and developed a CAD (Computer Aided Detection) based on neural networks, now available as cross check in the mammographic screening, as well as an integrated station used in training programs. The project directly involved several mammographic centres in Italy (Bari, Livorno, Napoli, Palermo, Sassari, Torino and Udine). It is very likely that the number of involved centres, as well as the database size, will rapidly increase.

However, the present network connection for these centres is not fast enough to allow the quick transfer of full images (about 60 MB each). For this reason, all the images are kept at a single site, where they are analyzed, and only a compressed image is transferred to the local screen requiring it. Indeed, the actual result is that each centre can fully exploit only its images. In fact, in order to set up a distributed database, made of the images digitized at all the different site, a Replica Manager Service is needed. In addition, given the limit set by the network connection, a tool to remotely analyze distributed data is required, as well.

The ALICE long term plan for the approach to the GRID phylosophy could be applied, with some modifications, to the management of distributed medical images on a shorter timescale and, obviously, on a smaller size; indeed, being based on the use of PROOF [8], it allows the transfer of the analysis code rathen than the input data (images). Clearly, if this approach proved to be effective, it could be suitable for any kind of input image.

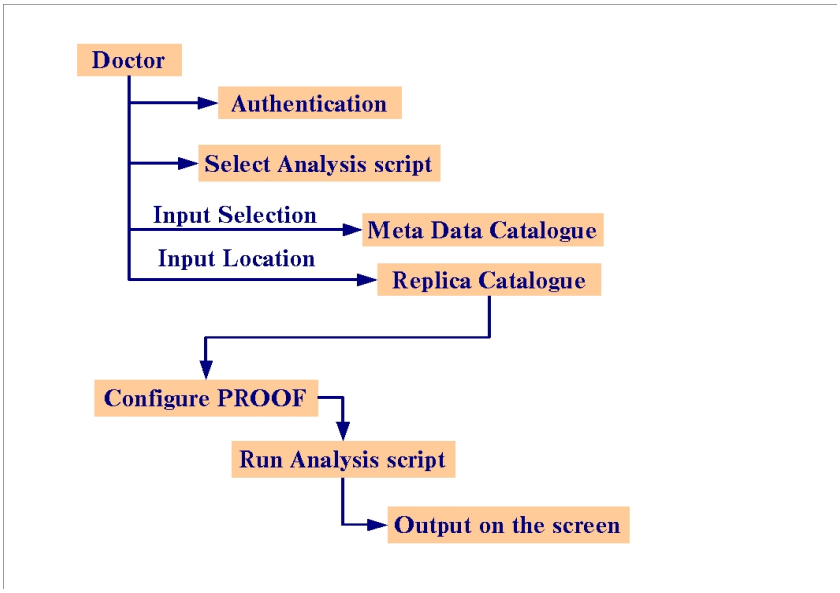

Figure 4: The logical flow for the analysis of images from the database

The scenario described above would therefore allow to keep the mammographic images where they are taken, and at the same time would make the full database available to all the centres, both for training and for analysis.

According to the requirements expressed in the report of the Teleradiology Working Group set up by the Medical Radiology Italian Society [12], there are three mainstreams to be followed in the near future:

- radiological teleconsulting;
- radiological telediagnosis;
- teletraining.

All of these activities are based on the sharing and analysis of remote images, by one or more people, regardless of their location. In this context, the GPCALMA (Grid Platform for CALMA) project was conceived to reach the following goals:

- CAD of mammographic images selecting particularly meaningful examples from the distributed database;
- study of the lesion size evolution as a function of time;
- telediagnostic screening with and without CAD;
- epidemiology, from of the radiological point of view, of breast cancers.

The two typical use cases for this application, corresponding to items 1 and 2 or to item 3 are showed in Fig. 4 and 5, respectively. Both are driven by the medical operator, via a Graphic User Interface as friendly as possible: the script to be run can be selected, as well as the input data. Then the Data Management Services are contacted and, according to the query results, the distributed system is configured and the script is processed, until the results are sent back to the submitter's screen.

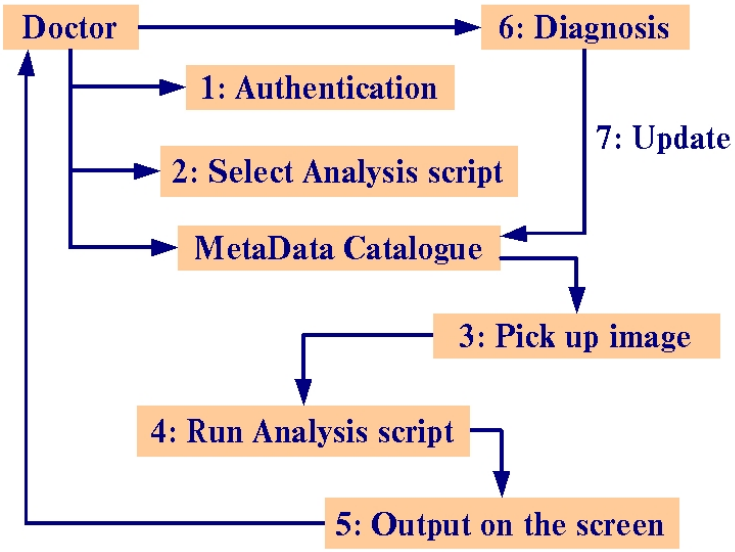

Figure 5: The expected logical flow for the telediagnosis.

In the case of telediagnosis, the final update of the database is also needed.

GPCALMA is now starting, and is proposed for a 3 years time scale, strictly related to the evolution of the EU−DataGrid project, even though results proving the effectiveness of the approach are expected to come earlier.

4. SUMMARY

Basic concepts related to the R&D in the field of GRID computing were discussed, trying to describe the main GRID Services and to sketch the overall outline of the system. A lot of activity is going on in the field all over the world, in the framework of several projects. A first prototype of a GRID system is expected to be released by the EU−DataGrid project in a few months. Many experimental groups, in HEP, Earth Observation, Biology and Medicine join the development, being potential GRID customers. In particular, two of these applications (the CERN/ALICE experiment and the INFN/GPCALMA project) were discussed in detail.

The setting up of an effective GRID system will most likely take a few years, but intermediate results are expected on a shorter time scale. The task is relevant and particularly difficult, but if successful it could represent an important improvement in the data and computing management for any geographycally distributed organization.

5. ACKNOWLEDGEMENTS

I am pleased to thank all my Collaborators whithin the EU−DataGrid, ALICE and GPCALMA projects, since most of the concepts addressed in this paper were progressively clarified by the continuous work and discussion involving all of us.

6. REFERENCES

[1] http://lhc.web.cern.ch/lhc
[2] http://www.eu−datagrid.org
[3] http://www.griphyn.org
[4] http://www.ppdg.net/
[5] http://alice.web.cern.ch/Alice
[6] http://AliSoft.cern.ch/offline
[7] http://www.to.infn.it/activities/experiments/alice−grid/
[8] http://root.cern.ch
[9] http://www.globus.org
[10] http://www.mrtg.org
[11] http://www.enginframe.com/sentinel/schema.xml
[12] "Indicazioni all'uso della TELERADIOLOGIA", report of the "Gruppo di Studio di Teleradiologia della Società Italiana di Radiologia Medica".