(PROPOSAL)

THE AGATA GRID COMPUTING MODEL FOR DATA MANAGEMENT AND DATA PROCESSING

version 0.6

July 2010
Revised January 2011

Mohammed Kaci[1] and Victor Méndez[1]

For the AGATA collaboration

[1]IFIC – Grid Computing and e-Science Group

[1] Instituto de Física Corpuscular, centro mixto Consejo Superior de Investigación Científica – Universitat de València, Valencia, Spain

Glossary and definitions :

AGATA : A European Project which aims developing, building and employing an Advanced GAmma Tracking Array spectrometer, for nuclear spectroscopy.

Raw Data : data collected from the experiment. The raw data contain the information about the line-shape of the signals delivered by the HPGe detectors.

Dataset : A set of files containing the Raw Data collected during one single run of the experiment by all the involved detectors.

Reprocessing : replay of the raw data by applying to them the Pulse Shape Analysis (PSA) and the γ-ray Tracking Algorithms.

Processing : refers to any of the data analysis and/or data reprocessing

Physics Data : reduced data, obtained after the reprocessing of the raw data (PSA+Tracking). The physics data are the ones that are used by the physicists to perform their analysis. They are of two types, the Intermediate Physics Data (IPD), obtained after PSA processing, and, the Final Physics Data (FPD), obtained after the Tracking processing.

IPD : Intermediate Physics Data
FPD : Final Physics Data

Official Data : refers to the Raw Data and the Physics Data. The Official Data are managed by The AGATA collaboration following an accorded Data Policy.

Data Policy : A set of rules and recommendations defined within the AGATA collaboration with regard to the use and the management of the Official Data.

GRID : Grid technologies allow that computers share trough Internet or other telecommunication networks not only information, but also computing power (Grid Computing) and storage capacity (Grid Data).

ACL : Access Control List
API : Application Programming Interface

BDII : Berkeley Database Information Index

CA : Certification Authority
CASTOR : CERN Advanced STORage manager
CE : Computing Element
CLI : Command Line Interface

dcap : dCache Access Protocol
DDM : Distributed Data Management

EGEE : Enabling Grids for E-sciencE
EGI : European Grid Initiative

FQAN : Full Qualified Attribute Name
FTS : File Transfer Service

GFAL : Grid File Access Library
GLUE : Grid Laboratory for a Uniform Environment
GPFS : General Parallel File System
GUID : Grid Unique Identifier

IS : Information Service

JDL : Job Description Language

LB : Logging and Bookkeeping
LCG : LHC Computing Grid
LDAP : Lightweight Directory Access Protocol
LFC : LCG File Catalogue
LFN : Logical File Name
LHC : Large Hadron Collider
LRMS : Local Resource Manager System

NGI : National Grid Initiative

PBS : Portable Batch System

RB : Resource Broker
RFIO : Remote File Input/Output
RFT : Reliable File Transfer

SE : Storage Element
SRM : Storage Resource Manager
SURL : Storage URL

UI : User Interface
URL : Uniform Resource Locator

VDT : Virtual Data Toolkit
VO : Virtual Organization
VOMS : Virtual Organisation Membership Service

WMS : Workload Management System
WN : Worker Node

CONTENTS

1. INTRODUCTION

The aim of the AGATA project **[1]** is to develop, build and employ an Advanced GAmma Tracking Array spectrometer, for nuclear spectroscopy. AGATA is being realized within a European collaboration and is intended to be employed in experimental campaigns at radioactive and stable beam facilities in Europe, over many years.

The AGATA collaboration consists presently in 44 institutes belonging to 13 countries, and around 400 members are involved in the project. Presently, in phase-1 of the project, the AGATA collaboration has constructed the demonstrator at the INFN-LNL laboratory (Legnaro – Italy) **[2],** which is running the first campaign of experiments. The demonstrator is made of five triple cluster detectors, each one being constructed with packing together three 36-fold electrically segmented High Purity Germanium (HPGe) crystals, of around 9 cm long by 8 in diameter each. In its future phases the AGATA γ-ray spectrometer will be moving to GSI (Germany) where 15 triple clusters are expected to be mounted and new campaigns of experiments will be performed during a couple of years. Then the AGATA ball, completed up to 60 triple cluster detectors (180 Ge crystals), will be mounted for another campaign of experiments (probably at GANIL).

At the present stage, phase-1 of the AGATA project, first experiments with 4 triple cluster detectors have been performed at INFN-LNL. The first data collection showed that, as the line-shape of the signals are recorded, the amount of raw data collected per experiment is very high compared with what usually obtained with the previous generation of γ-ray spectrometer arrays. It is of two orders of magnitude higher. About 40 Terabytes (TB) raw data have been recorded for the five commissioning experiments and more than 70 TB data for the last seven experiments. An average of 10 TB raw data per experiment have been recorded **[3]** with the demonstrator (up to 200 Gigabytes only per experiment obtained with Euroball). Moreover, these quantities of collected data have to be reprocessed using the Pulse Shape Analysis (PSA) and γ-ray Tracking algorithms, and this operation appeared to be very time consuming. The AGATA collaboration has then to face not only its data storage needs but also the reprocessing of such a huge amount of collected data.

The storage of such high amount of data as well as its reprocessing may drive the AGATA collaboration toward new computational technologies, different from that one used up to now in nuclear spectroscopy experiments. Particularly, when taking into account that the next phases of the AGATA project will deal with an increasing number of Ge crystals (AGATA 1π, AGATA 2π and full-AGATA), which would increase the computing and storage requirements.

The Large Hadron Collider (LHC) at CERN was developing Grid technologies and infrastructures for High Energy Physics experiments. Such project, called LHC Computing Grid (LCG) **[4]**, has been extended to other domains as diverse as Astronomy, Biomedicine, Computational Chemistry, Earth Science and Financial Simulations. These infrastructures are becoming these last years more stable and robust and could be a serious alternative solution for data storage, reprocessing, and data analysis for the AGATA collaboration.

A computational Grid **[5]** is a geographically distributed system aimed at putting together large set of heterogeneous computing resources (computing power, storage capacity, software applications, data etc.) among communities of users federated in a so-called Virtual Organizations (VOs). VOs are sets of individuals and institutions that agree upon common policies for sharing and accessing the computing resources. The interaction between users and the resources is made possible by a software layer known as middleware. It is a set of components that provides the user with high-level services for scheduling and running computational jobs, accessing and transferring data, obtaining information on the available resources, etc.

The AGATA collaboration involves a quite high number of people that share a common spectrometer array that will be moving with time across different experimental sites. The AGATA collaboration has already joined the Grid by creating its Virtual Organization (VO: **vo.agata.org**) in the year 2007 [6], and is using it (the Grid) by uploading the raw data obtained from the experiments to a Grid tape storage. However, no more extensive usage of the Grid resources and services is made, by the users, with regard to their needs in terms of performance of data processing and data management. Additional advantages in using the Grid are the followings:

- The event by event processing of the collected data makes this operation easily portable to Grid
- Free the data acquisition from suffering any additional dead time due to sophisticated online PSA reprocessing
- The experimental raw data are saved for future reprocessing
- The Physics Data are stored on the Grid
- The off-line data reprocessing may be performed as many times as necessary, with any new algorithms/versions, to produce the best quality physics data for the collaboration

Thanks to the worldwide LHC project, Grid infrastructures are presently installed yet and working at all the countries involved in the AGATA project. Moreover, these computing infrastructures are located at or near most of the institutes collaborating in AGATA. In addition, these countries are also involved in the "EGI-InSPIRE Project" **[7]** (following the EGEE project **[8]**) which is supporting, among other VOs, the VO vo.agata.org.

The AGATA collaboration may then use these infrastructures by contributing with computing resources or through agreements with these Grid sites.

This document presents a Grid Computing Model for the AGATA Data Management and Data Processing, designed with the objective of facilitating the data access, data transfer, data reprocessing and data analysis for the members of the AGATA collaboration, while reducing at its minimum the costs for that.

Section 2 presents the General Grid Infrastructure that would be used by the AGATA collaboration, without a big effort of development, as most of the needed Grid services are already operational. The Grid services and applications to be specifically adapted to AGATA are shown.

Section 3 describes the structure adopted for this AGATA Grid Computing Model. This structure follows the one currently used in the High Energy Physics (HEP) domain, but it is adapted to the specific needs and size of the AGATA collaboration. The AGATA Grid Computing Model is designed on the basis of the existing and well operating Grid infrastructures in HEP (Tier1s and Tier2s).

The Data Management and the Job Management are two key concepts when dealing with Grid computing. Sections 4 and 5 describe how these concepts are adapted to the AGATA collaboration, in the frame of the present Grid Computing Model.

Section 6 is dedicated to evaluate, within the present Grid Computing Model, the computing storage capacity and computing power required for AGATA. The evolution of these computing resources with time is also estimated.

Finally, section 7 deals with the support that would be provided to the members of the AGATA collaboration regarding the use of the Grid for data management, data reprocessing and data analysis.

2. THE AGATA GRID ARCHITECTURE

In this section, a general Grid architecture adopted for AGATA will be discussed and the different components involved will be described. Most of the components (resources and services) that AGATA will exploit are part of the general operation of the LHC Grid infrastructure. However, as discussed below, few components have to be specifically deployed and configured, and others developed, for AGATA.

The AGATA Grid architecture is based on the gLite middleware **[9]** but few cases also use suitable alternatives like Globus Toolkit services **[10]** or other third-party software. Currently, central core Grid services span all the areas of the gLite architecture, namely security, information system and accounting, data

management, and job management. Job management services include the Workload Management System (WMS, **[11]**), responsible for distributing and managing of tasks across Grid resources, and the Logging and Bookkeeping **[12]**, responsible for tracking WMS jobs as they are processed by individual Grid components. Information system services include several top level Berkeley Database Information Indexes (BDII, **[13]**), providing detailed information about Grid services which is needed for various different tasks. Security services include the Virtual Organization Membership Services (VOMS, **[14]**). A centralized LCG File Catalogue (LFC) **[15]** is being used as a data catalogue containing logical to physical file mapping.

Fig. 2.1 shows the AGATA Grid architecture needed for the adoption of a compromise between the presently available AGATA computing resources, the data storage and processing requirements of the project in the production stage, and the state of the art Grid technologies.

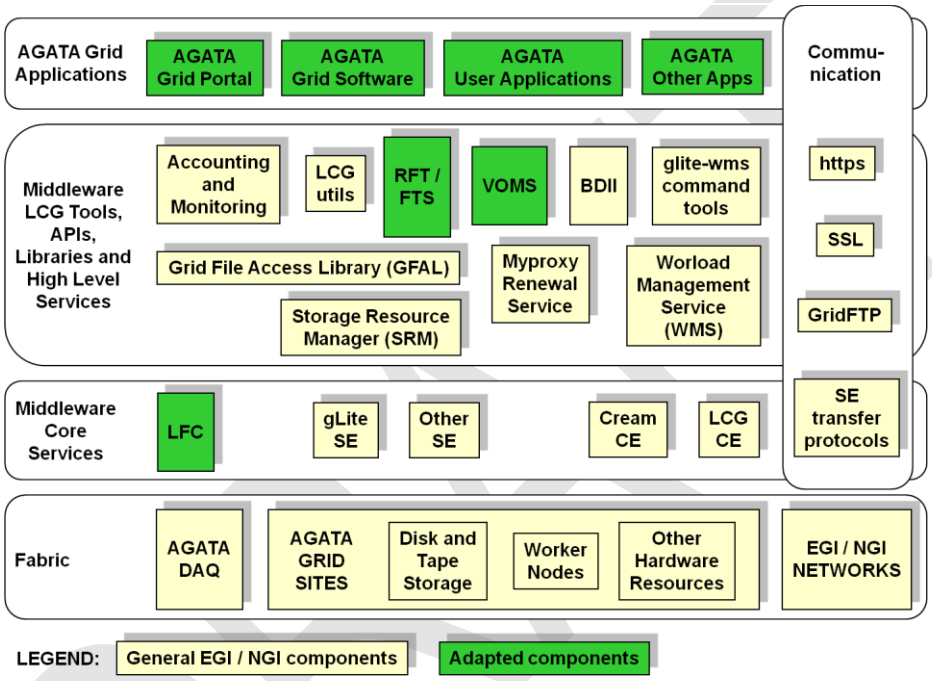

Figure 2.1 : The AGATA Grid Architecture. In green (dark) the components that have to be deployed, configured and developed specifically for AGATA.

Fig. 2.1 shows two types of architecture components. The general EGI / NGI components are generic components for many VOs (not only for **vo.agata.org**). For example the existing WMS will provide job management to the **vo.agata.org** just like to the rest of EGI VOs and does not need a special configuration or administration to support **vo.agata.org**. The presented architecture will use all the potential general services of the EGI/NGI infrastructures in order to reduce the simplicity of deployment and administration.

The AGATA project belongs to a European VO, **vo.agata.org**, that was approved in the EGEE framework, which now continued by the EGI-InSPIRE project. This means that any regional Grid infrastructure integrated in the EGI/NGI is nowadays giving some services to the **vo.agata.org**. Such services are basically an information system service (BDII), integrated in the EGI/NGI information system scheme of top-bdii and local-bdii; some Computing Elements (CEs) from IN2P3; some Storage Elements (SEs) on tapes and disk caching from CNAF/INFN; additional SE with tape and disk caching from CCIN2P3; and the user facilities associated to the VO members in their institutes, like the access to a User Interface (UI) or user disk space. Note that figure 2.1 is about services, not about servers and clients, which in Grid is so intricate while a server of certain service can be at the same time a client of another service, which is needed to give the first service.

The fabric layer of the AGATA Grid architecture is the AGATA DAQ, which is the Data Producer. The Grid Sites are used for data storage and data processing (PSA, γ-ray Tracking, Data Analysis). Thus, the available resources in the Grid sites can be Worker Nodes (WN) for Computing, different storage media, and other hardware resources, like computer machines to deploy the UIs to give the users access to the Grid. Other part of the fabric are all the remote networks available in the EGI/NGI context, which connect the Grid Sites.

Over the fabric layer there are three other layers:

- The middleware core services, which are an abstraction of the physical fabric computer resources, storage and computing, to "gridify" them.
- The middleware LCG tools, APIs, Libraries and High Level Services, that unify the use of the different core services in order to maintain the system integrity and coherence, and also, to have a single entry point for common operations that affect the different core services. The APIs of this layer are the client side APIs for the programming of the Grid applications.
- AGATA Grid Applications layer, that uses all the Grid middleware infrastructure.

The communication layer consists in transversal services that join the three other layers, without the classical software architecture, which communicate one layer with the neighbours. Thus, the application layer can access directly to a core service without the mediation of High Level Services layer. For example, an application can access to a specific SE using the corresponding SE transfer protocol, without the mediation of LCG tools, which maintain the coherence between the SEs and the catalogue (LFC). Of course, in these cases, the Grid integrity and coherence is the responsibility of the application or the user.

In addition to the generic EGI / NGI architecture components exploited by AGATA but that not need specific deployment and/or configuration, few Grid components have to be adapted to the specific needs of AGATA. Such services are the VOMS service and the LFC catalogue. Moreover, specific user oriented Grid applications for AGATA should be developed (AGATA Grid software, Grid tools,...).

2.1. The Secure Authentication and Authorization

The authentication and authorization are the standard EGI / NGI services and there is no need of special deployment for AGATA. However, **vo.agata.org** needs a specific configuration of gLite Access Crontol List (ACL) to the resources, with the proper rights and permissions for the "Groups" and "Roles" of **vo.agata.org** (see section 4.6)**.** From the deployment point of view this is affordable with the available EGI / NGI resources, just transferring the proposed ACL scheme to the VOMS server, and using the rest of EGI services.

Fig. 2.2 shows the general Security and Authentication model with the aim of informing about generic vo.agata.org user operations. The first step of a Grid user is to obtain a X509 certificate issued by a trusted Certification Authority (CA) (step 1). The certificate should be stored in a certain Grid system directory or/and in the browser. It represents the identified authentication of the user (steps 2 and 3). After that, any time the user wants to operate on the Grid, he has to request a temporal proxy (step 4). The obtained proxy define the VO the user belong to, and the associated VOMS attributes are the Full Qualified Attribute Name (FQAN) that will identify the grid group and role of the user inside the VO (steps 5 and 6). Proxy and FQAN have a temporal validity for security reasons, and they allow the user to access to the Grid services, mapping such attributes (FQAN) with some service rights and permissions.

Any time that a server is requested to give a service, the user proxy and VOMS attributes are sent in a secure way to the server (step 7). The grid server checks the proxy authentication with the same trusted CA certificate, which the proxy certificate was created with (step 8), and the attribute part is checked to get the active FQAN authorization from the VOMS server (steps 9 and 10). If the user is authorized by the VOMS server, then the Grid service performs the mapping with the local ACL to get the rights and permissions corresponding to this FQAN, and starts the requested service (step 11).

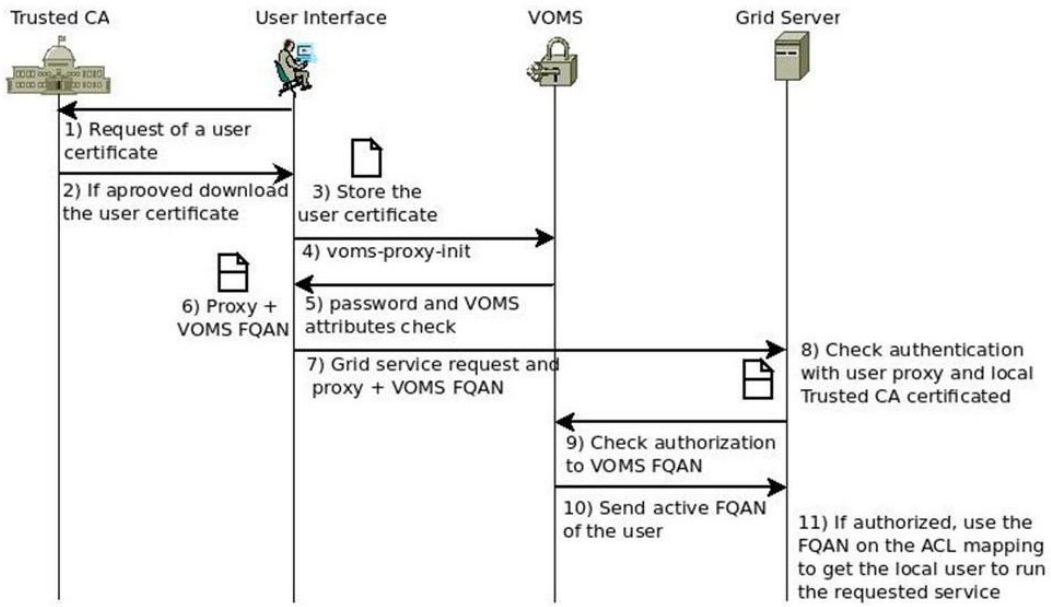

Figure 2.2 : Secure Authentication and Authorization model

2.2. The Information System and Monitoring

The Information System (BDII), and the Accounting and Monitoring Systems for **vo.agata.org** can use de EGI / NGI deployed services without any special configuration.

The Information System scheme is composed of a top-bdii and many local-bdii that collect site information to export to the top-bdii repository. The BDII provides mainly two classes of services, the LCG information to the users, that shows the available resources for a VO, using LCG-tools command line or LDAP queries; and the match-making system, which relates a specific resource request with the resource allocations that fulfil the request requirements in a specific moment. For this purpose, the information is structured using the GLUE Schema **[16]**. Anyway, the internals are transparent to **vo.agata.org**, since the most of operational issues are already supported by EGI / NGI.

However we have to take in consideration the maintenance of the **vo.agata.org** Information System coherence. Like in other VOs, there should be an administrator who has the responsibility of periodical pools the **vo.agata.org** Information System.

The Accounting System and the Monitoring System have no special interest for the AGATA users. These services are oriented to the Grid administrators convenience, and are fully integrated along the NGI regional monitoring and accounting systems, and the corresponding EGI global services.

The same as in the Information System case, the vo.agata.org should has a contact person to manage any issues coming from the EGI / NGI accounting and monitoring teams.

2.3. The Data Management

The AGATA project has a forecast of a huge tape storage requirement for Raw Data and also the challenge of the management of Physics Data in order to provide authorized access to the end user physicists of the

different experiments. The catalogue system is a key issue to have a coherent data management system. The catalogue service for gLite middleware is the LHC File Catalogue (LFC), which should be adapted for AGATA (see section 4).

A standalone LFC service will deal better with some performance issues than just adding the new **vo.agata.org** to any of the existing LFC services that at the same time are supporting other VOs, specially when the catalogue reaches an important amount of entries. The catalogue has to complaint the policy of rights and permissions for the different groups and roles. Such policy in vo.agata.org is defined in terms of ***rights*** over the files, who is the owner and the owner group of these files, in a unix like filesystem; also in the ***permissions*** to read, write and execute the files; using for his purpose the implementation of the ***group*** that a user belong to, that in **vo.agata.org** is the group of an experiment, and the ***role*** that the user has within the group. Some administrative operations are needed in the catalogue to maintain the system, actually, the configuration and administration operations are detailed if we compare with other VOs, which have public access to the data.

Another important service that needs to be adapted to AGATA, is an advanced data transfer service. The gLite has the File Transfer Service (FTS) **[17]**, that could be exploited by AGATA. An alternative to FTS is the Globus transfer service, called Reliable File Transfer (RFT) **[18]**, which is functional equivalent to FTS and good enough for the purposes in **vo.agata.org**. However, the integration of the Globus RFT in the gLite architecture, needs a special deployment, based on a VDT tools to deploy the Globus service in the Scientific Linux of the EGI, and also needs to be configured to work with the gLite services.

Other AGATA specific solutions for data transfer between the AGATA Grid sites may be investigated.

The rest of the involved data services, and other needed High Level Services for data management, can be taken as general EGI / NGI components with no special adaptation for vo.agata.org.

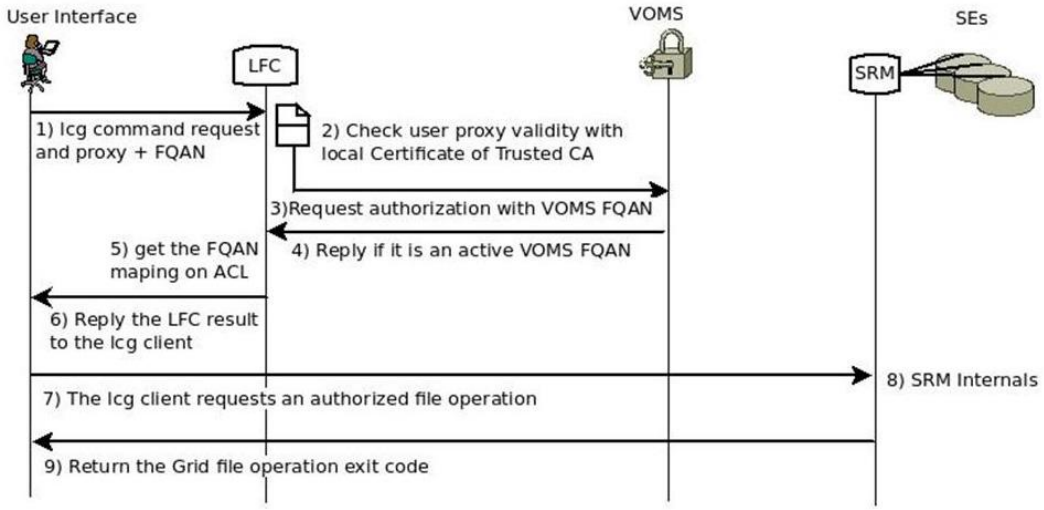

Figure 2.3 : A generic lcg command data management model

Fig. 2.3 shows the general Data Management clients and servers model for a generic lcg tools data operation. Some internal details are not shown, for example, the storage systems can be Castor, dCache, GPFS, or another available for **vo.agata.org** in the Grid Sites. Each of this storage systems has an access protocol on Grid, for example RFIO for Castor **[19]**, dcap for dCache or the POSIX/RFIO of stoRM **[20]** for GPFS. Such protocols can be accessed through the SRM Grid interface that unified the file Grid access. This is transparent to **vo.agata.org** since Grid Site administrators give this support to every EGI VO. Such SRM or access protocols

are also used to read the AGATA DAQ raw data, and they can be invoked by any file transfer service.

There are some considerations related to Fig. 2.3. The former is the double check on authentication and the double check on authorization. The first check was the Trusted CA certification and the VOMS proxy initialization, as it has been shown in Fig. 2.2. The second check is done by every Grid service like LFC, when the proxy is checked with the Trusted CA certification stored in the server, and the attributes of the proxy are remotely checked on the VOMS server (steps 2-4). It is the same we have seen above in the Secure Authentication and Authorization model in Fig. 2.3. Once the operation is authorized (step 5), the LFC perform the part of the lcg command on the catalogue and responses to the lcg client (step 6). The lcg tools data operation uses the LFC service and also the data services (steps 7 to 9) in a transparent way to the user. Lcg tools use the LFC and the GFAL libraries to integrate the catalogue and the Grid storage system.

2.4. The Workload Management

The Workload Management services for **vo.agata.org** can use the EGI/NGI deployed services in a transparent way for the AGATA collaboration purposes. The configuration and administration of such services can be supported by the EGI/NGI operation teams.

Fig. 2.4 shows the generic job management model. It is a brief description to comment some considerations for **vo.agata.org.**

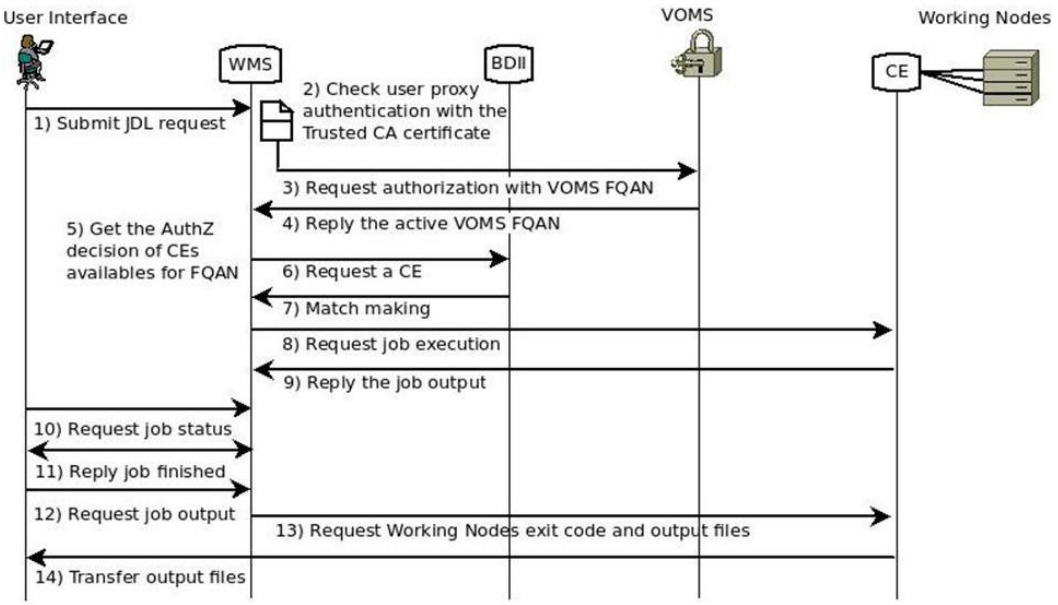

Figure 2.4 : Generic job management model

The user connected to a UI submits a workload in a Job Definition Language (JDL) **[21]** scheme (step 1). The authentication and authorization is done with local trusted CA certificate and BDII (steps 2-4). The Workload Management Service (WMS) uses the FQAN to gets the AuthZ decision for CE assignment, takes in consideration the VO-wide attribute and policy assertions, domain specific policy such as a set of VOs that has allowed access to a particular computational resource, and any locally defined site policy on the CEs (steps 5-7). The WMS submits the workload to the CE, and manages the job status (steps 8 and 9). The user has to manually

request the status of the workload to the WMS, and also request the job file outputs (steps 10-14).

Fig. 2.4 does not detail the CE management, which is an interface between a site and the Grid. It is transparent to the vo.agata.org since it is responsibility of the local sites administrators, who deploy the local job queues (like PBS…) and a CE service (LCG-CE or CREAM-CE), configuring the job policies according to FQAN (VOs, Groups and Roles).

This model is not enough user-friendly to vo.agata.org. There is an usability restriction because the user has to deal with JDL syntax. Other issue is the command line job management (steps 10,12) that can discourage an efficient use of Grid resources. For these reasons, it is necessary to use applications to simplified the job management of the user, as we show below in the the application layer subsection.

The generic job submission model of Fig. 2.4 can be applied for AGATA data reprocessing (PSA, γ-ray Tracking) and data analysis. To reduce the execution wall time, particularly for raw data reprocessing that uses data of high size, it is recommended to assign a WMS jobs to the CEs located at the same Grid site where the raw data are stored (see section 5).

2.5. The AGATA Grid Application Layer Considerations

The architecture description has argued some technical and functional requirements of advanced Grid applications for **vo.agata.org**. Fig. 2.1 shows four types of applications. The AGATA Grid software is discussed in section 5.2. The Grid portal is a Web graphical interface able to integrate the hold data management in a transparent way to the user. The AGATA user applications for data processing should be good enough for a user-friendly framework.

3. THE AGATA GRID COMPUTING STRUCTURE

The present Grid computing model for the AGATA project is valid as far as the line-shape of the signals delivered by the Ge detectors are recorded, and the PSA and γ-ray Tracking processing are performed off-line. As the line-shape of the delivered signals is recorded, the AGATA collaboration will deal with a high amount of collected data. The present Grid computing model is designed to reduce the cost of the computing resources used and the time needed to reprocess the data. Particularly, the storage space is reduced by sharing the available one, avoiding the unnecessary multiple storages, which is the case when each institute (or group of institutes) wants a copy of the raw data at home. In addition, storing the raw data and reprocessing them on the Grid will provide the end-users with the physics data necessary for their analysis within a reasonable period of time, depending of the size of data to reprocess (see section 5.4).

On the other hand, sharing a Grid storage will avoid also unnecessary data transfer of the raw data to the home institutes, and then save a non negligible time. Transferring 1 TB data between two sites at a throughput of 10 MB/s takes around one day. For an experiment which produces 10 TB in average one needs more than 10 days to transfer the data.

The data reprocessing is also a critical issue for the AGATA collaboration. It has been experienced that running the PSA is very time consuming (see section 5.4) as it is performed on local single computer, even with multiple cores **[22]**. On the other hand, the γ-ray Tracking processing is also expected to be time consuming, particularly for the next phases of AGATA where the number of crystals will increase.

Preliminary tests of running PSA and γ-ray Tracking using Grid resources has shown a considerable reduction of their processing time. This encourages the AGATA collaboration to use the Grid solution for data ***processing computing***. In addition, even the final data analysis could easily be adapted to Grid, which will avoid buying extra resources by the home institutes

Another point that has to be taken into account in this Grid computing model is that, within the AGATA collaboration, various experiments will be performed by different groups of users, which some times can overlap. A given user can be involved in various experiments with different groups of people. The Grid computing model provides the solutions to manage the access to the data in such a way that only a member of an experiment can access the corresponding data.

Taking into account all the above mentioned, the present Grid computing model for AGATA is based on the following ideas:

- All official AGATA data (raw data and physics data) are under the responsibility of the collaboration.
- The Grid resources dedicated to AGATA are managed by the AGATA collaboration.
- The DAQ farm must be used only for data acquisition and temporary storage for the most recent experimental raw data. It must not be used as a permanent storage or for reprocessing or analysis.
- The raw data collected from all the experiments must be distributed over the Grid sites following the policy (rules) described later in this document, avoiding unnecessary transfers of a huge amount of data.
- The raw data are stored, for security and integrity, on tape, in a redundant system. These data must not be directly accessed for reprocessing.
- For data reprocessing, a copy of the raw data must be available on disk. The data reprocessing is responsibility of the group (experiment). The data reprocessing is performed on the Grid.
- The produced physics data are distributed across the Grid sites for analysis.
- The physics data distributed on the Grid are organized in such a way that members of a group access only the data corresponding to their experiment.
- Data analysis with physics data is responsibility of the end-user physicists. They can use non-Grid or Grid resources.
- The end-users must be provided with a sufficient Grid disk space to store temporarily their output files when running analysis on the Grid.

The structure of the computing model for AGATA is inspired from the LHC computing model and follows the existing yet Tier1 and Tier2 Grid infrastructures. However, it is adapted to the needs and the size of the AGATA collaboration. The existing structure in Tiers is maintained, but it is worth noting that the present model does not make a strict frontier between these Tiers as all Grid sites in AGATA will operate similarly. In AGATA, a Tier1 site differs from a Tier2 one only because it provides tape storage, otherwise the Tier1 and Tier2 sites in AGATA operate in the same way. Consequently, the CPU and disk resources must be distributed, as much as possible, quite uniformly between all of the Tiers. Fig. 3.1 shows a representation of the AGATA Grid computing model, and the role of each kind of site is defined in the following sections:

3.1. The AGATA Data Production Site

The AGATA Data Production Site is the site where the experimental setup is located (γ-ray spectrometer array and DAQ farm). This site will be changing geographically its location as the AGATA spectrometer will be moving over different sites. The AGATA spectrometer is being installed and running successively at various experimental areas, where campaigns of experiments will be performed with different facilities and detectors. The campaign of experiments that is presently performed at INFN-LNL with the demonstrator (phase-1), will be followed by another one with AGATA phase-2 installed at GSI (Darmstadt – Germany). For the next phases AGATA will be moved to other sites for more experiments (GANIL – France,...).

The data taking is done at the AGATA Production Site. In the present Grid computing model for AGATA, the raw data obtained from an experiment is stored temporarily at the local disk storage on the DAQ farm and is transferred as soon as possible to the Tier1s data tape storages. In a normal operation of the Data Production site, a copy of the raw dataset that are completed (i.e. the run is done and its corresponding data files closed) starts immediately to be transferred to one of the Tier1 sites tape storage. Meanwhile, the data taking continue at the AGATA experiment for the following runs. The integrity of the transferred datasets must be checked. Once all

the experimental datasets (all runs) are transferred to the Tier1 and their integrity checked, they are removed from the AGATA DAQ farm to free disk space for the following experiments.

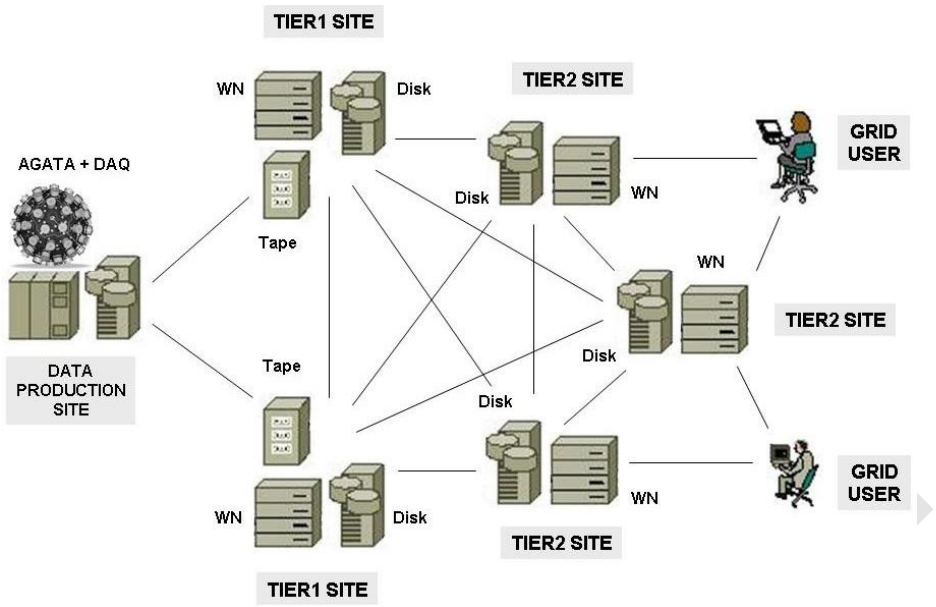

Figure 3.1 : The AGATA Grid Computing Structure

However, the disk storage of the DAQ farm must be able to store raw data equivalent to up to 3 experiments in order to face any unexpected problem that affect the data transfer to the Tier1 site. In case of problem (failing transfers, Tier1 temporarily unavailable, etc…) the raw data of a given experiment may be kept on the DAQ farm storage until the problem is solved (that should be done as quick as possible). But, sufficient disk space must be then available on the DAQ farm for the following one or two experiments. In any case, the operations between the Data Production site and the Tier1s must be coordinated in such a way that when a new planned experiment has to start, disk space at the DAQ farm must be already available and provided, to avoid delaying the planning of the experiments.

3.2. The Tier1 and Tier2 Sites

The AGATA Grid sites, namely Tier1s and Tier2s, are aimed for data storage and data processing. They must provide then a Grid infrastructure with sufficient disk storage capacity and computing power as well as the Grid services necessary for their operation. In addition to these computing resources, the Tier1s provide a tape storage system in order to store the raw data collected at the production site.

The raw data transferred from the Data Production site and arriving at a Tier1 site, are stored on tape as original data. The integrity of the data is checked during the transfer operation, which in case of failure would automatically trigger a repeated transfer. The transfer channels from the production site to all the Tier1s must be highly available. RFT offers an advanced management of the data transfer channels to ensure network bandwidth availability. In case that one of the Tier1 sites is failing, the data transfer is redirected to another available Tier1. Moreover, transfer channels must also be highly available between all the Tier1 sites in order to face any particular data redistribution.

The AGATA collaboration must provide more than one Tier1 site in order to secure the raw data and face any unexpected downtime of one of them during a campaign of experiments. For the security of the raw data, these latest are duplicated to another Tier1 tape storage, in order to create a redundant system from where data can be restored in case of problem. The complete raw data of a given experiment must be stored together on tape, avoiding the splitting of the data between two or more Tier1 sites.

The raw data stored on tape at the Tier1 sites are distributed across the Grid sites (Tier1s and Tier2s), and stored on disk for the PSA and γ-ray tracking processing. In order to reduce as much as possible transfers of a huge amount of raw data from a Tier1 tape storages to Tier1 and/or Tier2 disk storage for reprocessing, only a small sample of raw data is transferred. This data sample are used for the pre-reprocessing tasks, which consists in testing, adjusting and finalizing the reprocessing software (PSA+Tracking) for the current experiment. When the software is ready, the entire raw data corresponding to the experiment is copied on disk, at a given site, and all of the raw data are reprocessed.

The computing power (CPU) provided at the Tier1 and Tier2 sites allows reprocessing the raw data by running the PSA and γ-ray tracking processing. This allows distributing the data reprocessing tasks across several Grid sites avoiding the overload of a particular site. The off-line data reprocessing tasks are performed at the sites that are closer to the location of the used raw data. The produced final physics data, FPD, once validated, are distributed across the Tier1 and Tier2 sites, stored on disk, and their access provided to the physicists end-users for analysis. Then, the corresponding processed raw data may be deleted from disk, following the Data Policy described in section 4.4, in order to get back the disk storage space for the following experiments.

In addition to the CPU resources provided for reprocessing, the Tier1 and Tier2 sites in AGATA must also provide sufficient disk storage capacity in order to host the multiple copies of the FPD produced by the collaboration and distributed across the Grid. The AGATA official data stored on the Grid disks (at Tier1s and Tier2s), and belonging to the same experiment, may be split and distributed across various Grid sites, in order to optimize the space storage. As these data are managed by a catalogue (see section 4.3), their location is ensured.

3.3. The Home Institutes

To perform their analysis, the physicists at their home institutes must have access to the FPD of the experiments that are belonging to. These data being stored on the Grid, the end-user physicists must be provided with the necessary tools and services to access the Grid.

All of the institutes involved in the AGATA collaboration have a Grid infrastructure in their countries, and most of them in their institution. Providing then the AGATA end-user physicists with an account on a User Interface computer and a minimum of Grid services to access the Grid resources is feasible.

The physicists involved in an experiment may run directly their data analysis on the FPD stored on the Grid. However, if they prefer performing their data analysis at their home institute, they have to download from the Grid a copy of the FPD they will use. It is then responsibility of their home institute to provide them with the sufficient space storage required for the data.

4. THE AGATA DATA MANAGEMENT

The AGATA spectrometer is designed to be used, over many years, at various experimental areas and coupled to a number of ancillary detectors as well as other detection setups like RMS spectrometers. The AGATA collaboration will deal with a number of data files increasing with time, during the whole life of AGATA.

In addition, the members of the VO **vo.agata.org** will be organized in groups, each group consisting in the members that are involved in the same experiment and will then share the same experimental data. Consequently, the management of the official AGATA data within the collaboration must take into account this point.

All of this leads the AGATA collaboration defining a data management policy that regulates not only the use of the data within the collaboration but also the way of using the Grid resources.

The official data that the collaboration will deal with are mainly of two types, the Raw data and the Physics data (see section 4.1.). All of these data are organized in files that must be clearly identified within the Grid storages, easily accessible, and well separated by experiment. Moreover, the data location, and search across the Grid storages must be made easy for the end users. For that, rules for the names of the data files must be defined, standardized. In addition, the data must be well organized on the Grid by type and by experiment. A Grid catalogue must then be created in order to manage the complete official data files produced by the AGATA experiments.

4.1. The Data type

The main types of data that are managed by the AGATA collaboration are the following :

The Raw Data: The raw data are those obtained directly from the experiment. They contain the information about the line-shape of the signals delivered by the HPGe segmented detectors of AGATA as well as the information coming from additional detectors if any.

The Physics Data: The physics data are of two types. The intermediate physics data (IPD), are those produced by running only the PSA processing on the raw data. These IPD contain the information about the energy and positions of the interaction points of the γ-ray inside the HPGe crystals. They are afterwards used as input data for the γ-ray tracking processing (and merging with data coming from ancillary detector, if any) in order to produce the final physics data (FPD), which are the data that will be used by the end-user physicists for analysis. The AGATA collaboration has to deal with both types of physics data during the reprocessing tasks.

The User Data: The user data are those produced by the physicists (matrices, spectra, root trees, …) after running data analysis programs using the final physics data (FPD) as input. If the data analysis processing is performed on the Grid, the obtained user data may be stored on the Grid for sharing them with other members of the same group (experiment), or downloaded to a local computer for analysis with local tools.

4.2. The Data Storage Structure and Names

As the AGATA collaboration will deal with thousands of data files corresponding to different data types and various experiments, rules must be defined for the names of the data files in order to make the location and the search for the corresponding data across the Grid as easy as possible.

In the Grid storage space, the AGATA data are organized according to “experiment” and “type” of the data. The data files obtained from an experiment must be stored under a directory with a name describing the experiment. The data stored in such directory (**experiment-directory**) must be organized by type. Few requirements must be followed for the AGATA data nomenclature. At first, the names of the experiment-directory in the storage space (Grid SEs) and in the catalogue (LFC) should be in part mnemonic and it must be possible to generate them using an algorithm that covers all possible experimental cases of AGATA. In addition, the names generated must be unique within the collaboration and must not be reused once they have been defined and applied to a given experiment. Finally, all members of the collaboration must be aware of the nomenclature rules, and be able and willing to apply them.

The structure of an experiment-directory name consists of a number of fields that could have some semantic meaning, and are separated by a minus character, "-". The number of fields may be constant and their maximum length must be defined. The total length of the experiment-directory name must be reasonable and make the applications able to process it.

The additional special characters that may be used in the names of the experiment-directory files are the dot, ".", and the underscore, "_". They may be used to separate parts of the same field. It is important that the experiment-directory name must be case sensitive. An experiment-directory name could have the following structure :

<Site>-<DateTime>-<ExpParams>-<Setup>/

where each field is defined as follows:

<Site> is the name of the site where the experiment is taking places. For example LNL, GSI, ...

<DateTime> is date/time of the experiment. It can be expressed for example as presently used in AGATA by "year_week" or also as "day_month_year" or whatever else.

<ExpParams> describes the parameters of the experiment. It could show the reaction used, the beam energy, the target thickness, the nucleus of interest. Any combination of these information could be used. Some self-explanatory expression may be used. Examples could be "32S110Pd_130MeV_0.3" or "32S110Pd_130MeV".

<Setup> describes the experimental setup used. For exemple "agt" for only AGATA, "agtnw" for AGATA+Neutron wall, "agtdante" for AGATA+Dante, etc…

The following are examples of correct names for experiment-directories :

LNL-2009.week43-32S110Pd_130MeV-agtLaBr3Si/
LNL-2009_week43-32S.130MeV_110Pd-agtLaBr3Si/
LNL-Oct_2009-32S.130MeV_110Pd_0.3mg-agt6Ge_LaBr3Si/

An experiment-directory is associated to a given experiment. It is a parent directory for the whole official AGATA data related to that experiment. It will contain sub-directories organized by data type. A directory structure for the storage of the data corresponding to an AGATA experiment could be as follows:

<exper_dir>/<data_type>/<run-number>/<detector>/<data_file>

Where,

<exper_dir> is the main directory for a given experiment, as it is defined above,
<data_type> is the type of the data, and must take one of the following value name : "rawdata", "IPD", "FPD",
<run-number> defines the run-number and could be for example: Run009, Run025, Run170, ...
<detector> defines the detector that fired, for example: 1B, 4G, Ancillary, ...
<data_file> is the name of the data file. Because the data file is located in a specific place in the directory structure, its name could be a simple string, like "event_mezzdata.bdat" as presently used in AGATA. And the same name may be used for different runs and detectors. However, to make the data file name unique (for any others purposes, like copying various data files in the same user directory), one can use a data file name structure similar to that used for the experiment-directory and add additional fields to inform about the type of the data, the detector fired and the corresponding run number. A data file name structure could then be :

<Site>-<DateTime>-<ExpParams>-<Setup>-<data_type>-<detector>-<run-number>-.dat

Where the various fields of the file name are defined as previously shown above in this section. Doing so, the following file name are correct for single data files obtained in an AGATA experiment:

LNL-2009.week43-32S110Pd_130MeV-agtLaBr3Si-raw-Run019-1B-.dat
LNL-2009_week43-32S.130MeV_110Pd-agtLaBr3Si-FPD-Run078-IPD-Anc-.dat
LNL-Oct_2009-32S.130MeV_110Pd_0.3mg-agt6Ge_LaBr3Si-FPD-Run165-.dat

Note: as the FPD are obtained after the γ-ray tracking and merging, their file name does not need the field <detector>, which is then skipped.

4.3. The File Catalogue

A file in the Grid is identified by its Grid Unique Identifier (GUID). A GUID is assigned the first time the file is registered in the Grid, and it consists in a string of characters obtained by a combination of MAC address and a timestamp to ensure its unicity.

guid:<36_bytes_unique_string>

However, a user will normally use a more friendly and mnemonic file names. Any AGATA data file in the Grid can be referred to mainly by its Storage URL (SURL) name and/or by its Logical File Name (LFN). The LFN identifies a file independently of its location while the SURL contains information about where the physical file is located, and how it can be accessed.

In a Grid environment, files can have replicas at many different sites, and all replicas must be consistent. This makes that Grid files cannot be modified after creation. They only can be read and deleted. Users do not need to know where a file is located, as they use LFNs for the files that the Data Management services use to locate and access them. All of the replicas of a given file have the same GUID but different SURLs and LFNs.

Users and applications need to locate files or replicas on the Grid. The mapping between GUID, SURLs and LFNs are kept in a service called a File Catalogue. It is important to mention that although data files are held in SEs, they are made available through File Catalogues, which record information for each file including the locations of its replicas (if any). The File Catalogue adopted by gLite is the LCG File Catalogue (LFC). This tool allows users to view the entire Grid as a single logical storage device. The mapping between the GUID, SURLs and LFNs file identifiers are shown in Fig. 4.1.

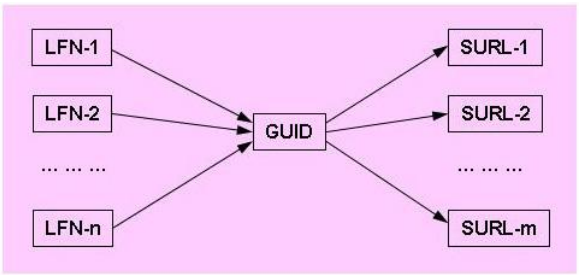

Figure 4.1 : mapping filenames in LFC catalogue

The only file catalogue officially supported in WLCG/EGEE, in the gLite middleware, is the LCG File Catalogue (LFC), which would be adopted by the AGATA collaboration. In the LFC, the LFNs are organized in a hierarchical directory-like structure and will have the following format :

lfn:/grid/vo.agata.org/<data_directory>/<data_file>

For the AGATA data, the structure adopted for the catalogue, depending of the data type, could be as follows:

lfn:/grid/vo.agata.org/<exper_dir>/<data_type>/<run-number>/<detector>/<data_file>

The SURL, which identifies a replica in an SRM-managed SE, could be for AGATA of the form below.

srm://<SE_hostname>/<vo-agata-data>/<exper_dir>/<data_type>/<run-number>/ \
<detector>/<data_file>

The AGATA collaboration must have a central catalogue to manage the organization of the data. The coherence and consistency of the data within the AGATA collaboration are maintained by the Grid catalogue (LFC). The catalogue allows the end-users finding easily the data they are looking for on the Grid. The LFC catalogue handles also the ACLs permissions through the group membership of the end-user and the role given to the end-users in order to organize the permissions access to the data.

This catalogue must be hosted in one of the Grid sites. It is possible to deploy read only catalogues to ensure the accessibility and improve the LFC performance. Another way to improve the LFC performance is using a read/write LFC standalone service giving support only to the VO **vo.agata.org [23]**. The LFC performance is a key issue in final application performances when the catalogue reaches huge size of entries.

Important note: a file is considered to be a Grid file if it is both physically present in a SE and registered in the file catalogue. In general, several high level tools like lcg_utils will ensure consistency between files in the SEs and entries in the file catalogue. However, usage of low level Data Management tools could create inconsistencies between SEs physical files and catalogue entries resulting in corruption of Grid files. This is why the usage of low level tools is strongly discouraged.

Note: All the official AGATA data stored on the Grid must be registered in the catalogue in order to maintain the coherence and the consistency between the replicas. Users ***must*** exclusively use the lcg-utils and the lfc-utils command lines in order to maintain this consistency.

4.4. The Data Management Policy

The official data produced by AGATA and their movement are under the responsibility of the AGATA collaboration that has to decide and apply the data policy accorded within the collaboration. This policy concerns the data management (transfer, replica, access, deletion) and storage. It concerns also the permissions granted for the data access, as well as the responsibilities of the users regarding their use of the data.

The Grid data management policy defined within the AGATA collaboration aims facilitating the data exploitation by users while improving the performance in using the Grid computing resources, by reducing particularly a lot of expected chaotic use of Grid storage space.

An AGATA Grid Computing Committee (AGCC) would be created to act, under the supervision of the AGATA collaboration, for what concerns the management of the official data on the Grid. Important decisions regarding the official data must however be taken by the AGATA collaboration.

Firstly, during the AGATA operation, the most consuming computing resources are the raw data that could need a lot of space to store them and high number of CPU cores to reprocess them, particularly for experiments that produce high amount of data. In order to optimize the use of the Grid resources, only a small but sufficient sample of the raw data of a given experiment is copied from tape to a disk space at a Grid site (Tier1 or Tier2). This data sample is used for previous testing and improving the reprocessing programs (PSA and γ-ray Tracking)

before running the final reprocessing of the whole experimental data.

The off-line reprocessing of the raw data must start as soon as the raw data samples are available on the Grid disk storage. This reprocessing is expected to be repeated several times, until the best quality FPD are produced. In order to avoid multiple work and unnecessary use of CPU time during reprocessing, it is recommended that only few members of the group perform this task. Once the reprocessing programs are finalized and validated, a copy of the whole raw data of the experiment is replicated to a Grid disk and the reprocessing is performed for the whole experiment.

The produced FPD must be validated (Data Quality Policy) before their distribution to all of the members of the group. If these FPD are popular (many members use them on the Grid for their analysis), multiple replicas of them can be distributed across various Grid sites. If not, only one copy is provided (may be distributed over various sites).

Once the reprocessing is completed, the raw data are removed from the disk storage to get back storage space for next experiments.

The FPD are considered as permanent data, stored on the Grid for all of the time they are exploited by their owner group members, which could extend to few years. Afterwards, the old FPD may be stored definitely on tape at the Tier1 sites and removed from the Grid disk storage..

The data reprocessing task, the validation of the FPD and the deletion of these data are responsibility of the spokesperson of the experiment. In case of more than one spokesperson for the same experiment, they should designate a responsible for these tasks, who will also act as a contact person for the group.

The user data disk is considered as a scratch disk and is cleaned periodically, removing first the old files.

In addition, to the few described data management policy, important rules that could be adopted for the data management policy, in the framework of the present AGATA Grid Computing Model, are listed in the following.

- The management of the official data is responsibility of the AGATA collaboration (or the AGCC).
- All the official AGATA data (raw, IPD, FPD) must be stored on the Grid.
- The read access to data is organized in such a way that members of a group can access only the data of their experiments. Agreement between groups for sharing data is responsibility of the groups.
- The write access to tape is restricted to only few members of AGATA. The write permissions are granted to these members by decision of the AGATA collaboration (or the AGCC).
- The read access to tape is granted to only few members named by the AGCC.
- The access to the data on tape is restricted to only the replication operations.
- No data processing is performed directly using the data on tape. All data reprocessing and data analysis must be performed using the data replicated on disk. The progress of the data reprocessing must be reported periodically to the AGCC in order to help in managing the disk space.
- Any data deletion from disk of official data must be authorized by the AGATA collaboration. It is executed by authorized member and supervised by the AGCC. The authorized members must then be granted with the write access to the disk space where the official data are stored.
- The replication of data on disk (disk → disk) is decided by the AGCC and executed by authorized members.
- The write access to disk where the official data are stored is granted to few members and decided by the AGCC.
- The reprocessing must be organized by the designated responsible of the experiment (one responsible by experiment and by institute) in order to optimize the use of the resources.
- The write access is granted to all users of the VO vo.agata.org on the dedicated users data disk storage
- Each operation of data replica must be reported in the catalogue in order to ensure the coherence and consistency of the data.

In the following sections are discussed the ideas related with the AGATA data management, taking into account the above described data management policy.

4.5. The Data Distribution and Placement

The raw data are produced originally at the Data Production Site. For each experiment, two copies of the raw data are stored on tape at two different Tier1 sites, creating then a redundant Grid storage that secures the data.

Fig. 4.2 shows the flow of the raw data within the AGATA Grid data storages. Following the data management policy, a small sample of raw data is replicated to a disk storage at the Tier1 and Tier2 sites in order to prepare and finalize the reprocessing software. Once the PSA+Tracking parameters are optimized for the current experiment, the entire raw data is replicated to disk storage and the reprocessing is performed.

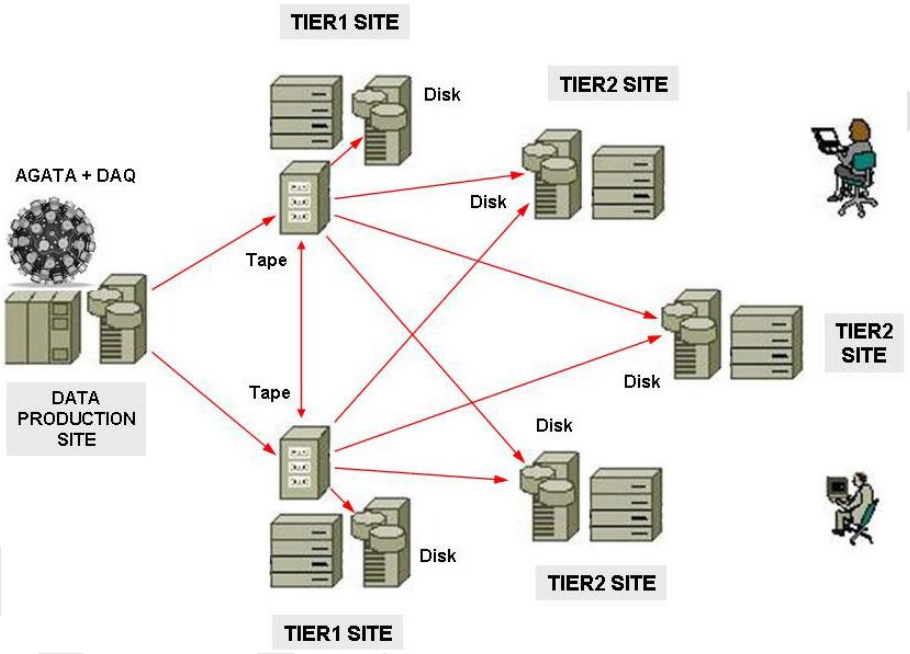

Figure 4.2 : Data Model: flow of the raw data

Note that it is not mandatory that the raw data of a given experiment be copied to the same Grid site. They may very well be distributed across more than one site, as they are well localized by the catalogue. In that way, the reprocessing is also distributed. This may increase the efficiency and performance of the reprocessing. When the data reprocessing is completed, this copy of raw data is removed from disk in order to free space for the next planned experiments.

The raw data reprocessing is performed, in general, in two steps. In the first step, the data reprocessing program reads the raw data, process the PSA and produces the intermediate physics data (IPD), which contain the energy and the interaction points of the γ-rays with Ge crystals. The second step of the data reprocessing, which could be more or less complex, depending on the experiment, consists in the γ-ray tracking processing which reads the IPD as input data and produces the final physics data, FPD. These FPD contain the information about the reconstructed γ-ray and are the data that will be used for physics analysis.

When produced, the IPD are placed in a dedicated directory on the SE. Once the whole reprocessing completed and the FPD produced, validated, and their access provided to the end-user members of the group, the IPD may be deleted from the Grid disk storage.

The FPD, as they are the most popular data, because used by all of the end-user physicists, must be distributed as uniformly as possible on the disk space available at the Tier1 and Tier2 sites. Fig. 4.3 shows what could be the flow of the FPD between the disk storages of the AGATA Grid sites. For security reasons, two copies of the final physics data, for each experiment, are stored permanently on Grid and distributed across the Tier1 and Tier2 sites. If a given FPD are popular (i.e. many users use these data for analysis) on the Grid, more replicas may be requested and provided during the data analysis phase. Multiple replicas of the FPD on the Grid ensure the high availability of the data (even if sometimes a given Grid site is unavailable) and improves the performance.

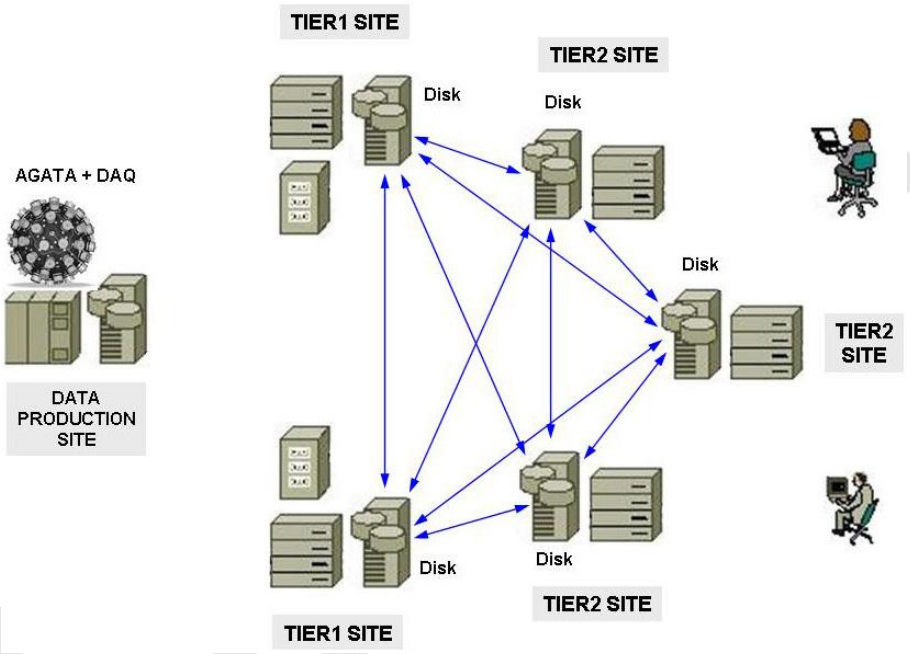

Figure 4.3 : Data Model: flow of the physics data

The end-user physicists can use the Grid to process the FPD for their analysis. They may also download the FPD to their home computers for local analysis, but then they should provide by themselves the necessary space storage for that. In addition, they must have installed the necessary Grid tools to do that. Moreover, if the FPD are not used, on the Grid, by their owners then only one copy is kept permanently on the Grid in order to save the data.

The end-user physicists will also produce their output files (spectra, matrices, root trees,…) from their data analysis performed on the Grid using FPD. These files may be produced directly on the dedicated users data Grid disk storage, then downloaded to their home computer, or, produced directly on their used UI computer.

The AGATA data files must be handled in the SEs with a mnemonic and logical structure. They must be organized by experiment and by type of data. It is recommended that the disk space at each AGATA Grid site be organized in directories that well define the experiments and the types of data. The following data storage structure format could be defined:

srm://<SE-hostname>/<vo-agata-data>/<exper_dir>/rawdata/<run-id>/<jB|jG|jR>/<datafiles>
srm://<SE-hostname>/<vo-agata-data>/<exper_dir>/IPD/<run-id>/<jB|jG|jR>/<datafiles>
srm://<SE-hostname>/<vo-agata-data>/<exper_dir>/FPD/<datafiles-run-id>
srm://<SE-hostname>/<vo-agata-data>/<exper_dir>/userdata/<username>/<user-files>

From the point of view of the catalogue, the following data format structure is derived from the above:

/grid/vo.agata.org/<exper_dir>/rawdata/<run-id>/<jB|jG|jR>/<datafiles>
/grid/vo.agata.org/<exper_dir>/IPD/<run-id>/<jB|jG|jR>/<datafiles>
/grid/vo.agata.org/<exper_dir>/FPD/<datafiles-run-id>
/grid/vo.agata.org/<exper_dir>/userdata/<username>/<user-files>

4.6. The Data Access

From the previously mentioned data management policy (see section 4.4), permissions must be well organized with regard to their data access by the users. Both the VOMS authorization service and the LFC catalogue manage the rights and permissions of the directories and file access of groups and users. The rights are the user ownership of the directory or file, and also the group ownership. The permissions are very much like those of a UNIX file system: read (r), write (w) and execute (x). A combination of these permissions can be associated to these entities:

A user (user)
A group (group)
Any other user (other)
The maximum permissions granted to specific users or groups (mask).
The Access Control List (ACLs) with the eventual permissions to other users and groups different than the mentioned owner user and group the user belongs to.

Permissions for multiple users and groups can be defined. If this is the case, a mask must be defined and the "effective" permissions are the logical AND of the user or group permissions and the mask.

In LFC, users and groups are internally identified as numerical virtual uids and virtual gids, which are virtual in the sense that they exist only in the LFC namespace, they correspond to the local users of the server in which the LFC is deployed. Such scheme is used in Grid services to do the mapping between the FQAN of a Grid user and the corresponding local user with the proper rights and permissions for the given FQAN.

In addition, like in unix like filesystem, a LFC directory inherits the rights and permissions from the parent directory, and they can be changed by the owner of the directory.

The general protocol of the authorization was shown in section 3 of AGATA Grid architecture, following we describe some internals of authorization method, to explain the proposed scheme and provide administration templates for Grid user management..

As was explained in section 2.1, any Grid user has a certificate that authenticates the identity of the user. Additionally, the VOMS server has a FQAN for each Grid user certificate, and this FQAN or attribute is attached to the user certificate in a temporal proxy, which it would be used for authentication and authorization of any Grid service, in our case the LFC service integrated in the Grid file system.

The FQAN defined for the current data model are the following :

GROUP = experiment with the nomenclature <year_weekN> (i.e.: "2009_week43"),
ROLE = "datamaster": the contact person and administrator of the AGATA Grid Data Management,

ROLE = "qualified": is the role for the users who can run both raw data reprocessing and data analysis,
ROLE = "common": is the role for the users who can only run data analysis (using FPD).

A VO-Admin Role is also defined and corresponds to the administrators of the AGATA VOMS server.

Such FQAN of the users voms proxy are used in the mapping between the Grid user authorization in a specific moment, and the local user of the Grid service, in our case LFC, that realize the corresponding rights and permissions. The mapping is performed with an ACL of the LFC service. The ACL defines the user categories that must be accepted by the LFC service provided by a site. It indicates for each category to which kind of local accounts the user should be mapped, where applicable.

Another important data access issue is related to the SE rights and permissions. While a replica can be accessed using low level utilities, the site administrators where the AGATA files are stored, must use the same authorization templates showed below, to mapping user authorization of the SRM service.

4.7. The Data Deletion

Once the data reprocessing completed and the FPD produced and validated, the contact person of the experiment notify the AGCC and the used raw data are declared erasable from the disk space. The deletion of the raw data is then decided by the AGCC after notifying the AGATA collaboration. When the deletion of the raw data is confirmed, it is executed by an authorized member of the group or by the system administrator of the site.

Similarly, the same policy is followed for the deletion of the FPD (if it proceeds). If the FPD corresponding to a given experiment, and stored on the Grid, are not used because their owners have chosen to previously downloaded them to their home institutes for local analysis, the multiple replicas of these FPD, if any, are deleted from the disk storage, keeping only one single copy of the data on the Grid. In any case, once the FPD have been already intensively analyzed a copy is stored on tape at a Tier1 site and all replicas on the Grid are deleted.

In case of urgent need of disk space, the first data to be deleted are the older raw data that have been kept on disk for very long time, and then the IPD and then the replicas of FPD. However such a situation would not happen in normal an careful operation of AGATA.

In case that members of the AGATA collaboration would like to reprocess raw data or analyse FPD, of old experiments, which have been already removed from the Grid disk storage, they must introduce a request to the AGCC for a new copy on disk from the original data stored at the Tier1 tape storage.

4.8. The Data transfer

Data transfer between AGATA Grid sites may be performed using one of the Grid data transfer services (FTS, RFT, lcg-utils). The File Transfer Service (FTS) is the one used by the gLite middleware. The user schedules the file transfer from source to destination while the sites control the network usage. The FTS handles internally the SRM negotiation between the source and the destination SEs. It is recommended to perform copies of sets of files instead of individual files. Data transfer operation includes a validation that verifies the integrity of the transferred files

The FTS is provided in the EU Grid infrastructure and AGATA can exploit its services, as it is doing with other common Grid services (see section 2). Presently, the main data transfer in AGATA concerns the upload of the raw data from the Data Production Site (experiment) to the Tier1s Grid sites using basically the lcg-commands.

In the framework of the present Grid Computing Model, and in relation with the data transfer between the AGATA Grid sites, a specific tool using the lcg-commands in a distributed environment including the interaction with a dedicated AGATA LFC catalogue is being developed and tested for AGATA.

Data transfer between sites must be reduced as much as possible, particularly for the high amount of data. Data transfer of a huge amount of data is performed by authorized members and for particular cases which could be the following:

- transfer of raw data from the Data Production site to Tier1 tape storage
- transfer of raw data from Tier1 tape to Tier1 and Tier2 disks storage
- replicate FPD between the Tier1 and Tier2 disks storage

During their data analysis, the users must avoid as much as they can unnecessary file transfers by submitting their data processing jobs to CEs that are located at the same sites where the input data are.

5. THE AGATA DATA PROCESSING AND JOBS MANAGEMENT

Running data reprocessing and data analysis on the Grid require a careful adaptation of the whole AGATA software in order to reach the best performances and optimize the use of the Grid computing resources.

In addition to the AGATA data processing, simulation calculations (Monte Carlo simulations and/or PSA calculations) can also run on the Grid if necessary.

5.1. gLite and the Job Management

The gLite middleware is based on the concept of "job submission". Fig. 5.1 shows a job flow in the gLite Grid middleware, as well as the job status at each stage of its life.

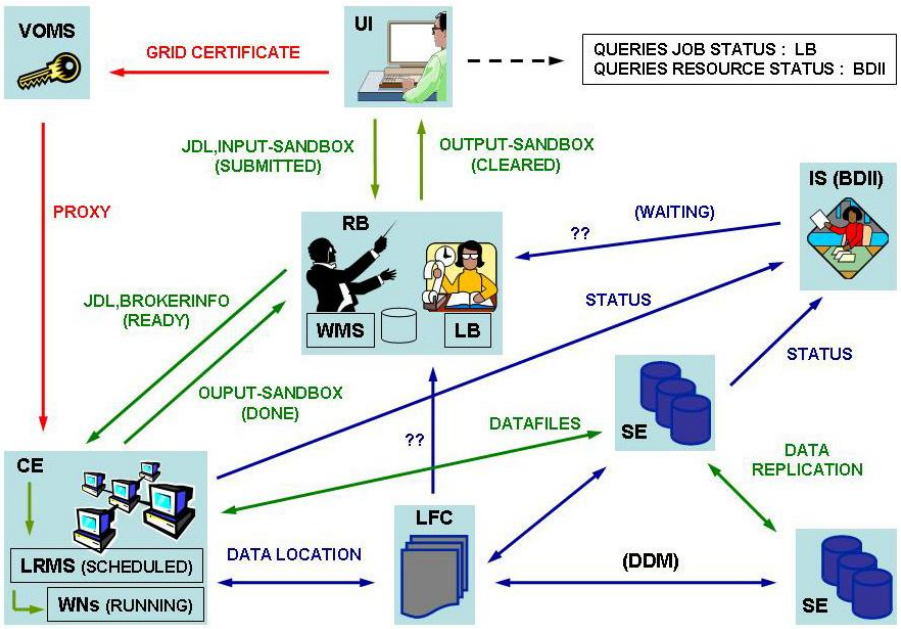

Figure 5.1 : Job flow in gLite middleware

A job is an entity that contains information about all the stuff needed for a remote execution of an application, such as its executable, the environment settings, the input data and the expected output files to be retrieved. All these parameters are defined in a text-file written in the Job Description Language (JDL) syntax **[21]**, which is a high-level specification language based on the Classified Advertisement language **[24]**.

End users can access the Grid through a Grid component called User Interface (UI), which is a client properly configured to authenticate and authorize the user to use the Grid resources. When a user wants to execute a computational job, he composes a JDL file and submits it to the Workload Management System (WMS), which is the service in charge of distributing and managing tasks across the computing and storage resources. The WMS, basically, receives requests of job execution from a client, finds the required appropriate resources and then dispatches and follows the job until completion, handling failure whenever possible.

The jobs are sent to Computing Elements (CEs), which manage the execution queue of the Worker Nodes (WNs), the computers where a job actually run. In other words, a CE acts as an interface to the computing farm installed in a Grid site.

Once the task has been completed, all the output files listed in the JDL are packed in the Output Sandbox and sent back to the WMS. Finally, the user can retrieve the Output Sandbox onto his/her UI. Applications that need to manage large data files (either as input or output) can store/retrieve the data by accessing directly the Storage Elements (SEs) using the data management API provided by the middleware.

5.2. Adapting the AGATA Data Processing Software to the Grid

The AGATA collaboration needs adapting its data processing software to run on the Grid. This process may comprises the creation of additional bash scripts as well as changes in the original source codes in order to include the needed APIs (Application Program Interface) to directly interact with the Grid components. In some case, it is also necessary to stop using services and libraries that are not supported by the standard distribution of the adopted Grid middleware.

In what concerns AGATA, the data processing software, including the PSA+Tracking reprocessing as well as any data analysis processing, must be adapted to Grid in such a way that optimizes the use of the computing Grid resources.

In order to run on the Grid, any AGATA software has to interact with other Grid components and services, particularly with the SE where the input data files and the output files will be hosted. For that, and depending of the file system used at each Grid site, various data access protocols would be used (dcap for dCache, rfio for Castor). These protocols are managed within some components of the GFAL library, which allow emulating a POSIX-like access to data. It is worth noting that Lustre file system for example do not need the GFAL components to deal with data access as it includes yet a POSIX data access.

The other point to take into account it that running the AGATA reprocessing software on the Grid is performed trough submission of Grid jobs to various WN. Consequently, the user must built correctly these jobs, including the JDL and the script file. The JDL file will contain all the information about the executable software, the input files, the output files and the specific requirements (if any) for the job execution. The script file will contain all of the actions that have to be executed remotely on the WN to perform the execution of the software and the production and retrieving of the output files.

One way to run any AGATA software (data reprocessing or data analysis) on the Grid is as follows. At first, a Grid job is defined that will contain a compressed copy of the software to execute together with its configuration files. It will contain also the necessary JDL and script files. Once the submitted job is delivered to a CE and starts running on a WN, the script file is launched. The package software is then unpacked, the software is compiled and then starts running. Of course, a precompiled version of the software could be

submitted to the Grid, in case of compatibility of operating systems.

However, in normal operation, all the necessary common software for the data processing must be tested by the collaboration. Afterwards, it is installed in the WNs of all of the AGATA Grid sites and then is tested once again on the Grid and validated. If several versions and/or algorithms are used, they must be well organized and identified to facilitate their use by the end-users.

The developed AGATA software for data reprocessing on the Grid has to read raw data as input files and produces FPD as output files. The data structure of the AGATA raw data, for a given experiment, is presently represented as a tree structure. The raw data are organized in a number of datasets, each dataset corresponding to one run of the experiment, and contains as much as data files as the number of the current Ge crystals used in AGATA. In other words, a dataset consists of a series of data files, each one containing the data collected by one single HPGe crystal (identified by nB|nG|nR) during the same run. The raw data corresponding to a given experiment is then a collection of datasets. Fig. 5.2 shows the organization of the raw data as currently stored at INFN-Tier1 (CNAF-Bologna, Italy) for Demonstrator.

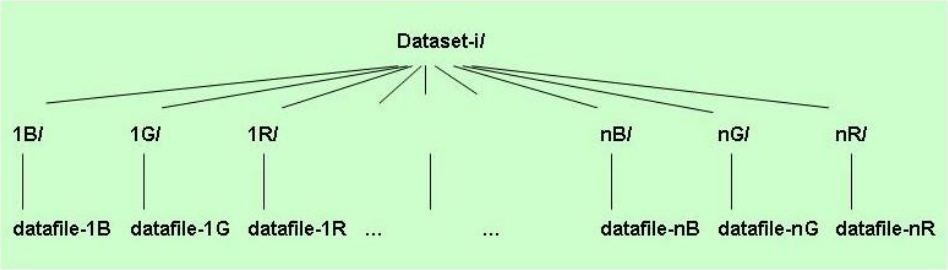

Figure 5.2 : Data structure, as written on the LCG-Tier1 INFN-CNAF storage system

For the PSA reprocessing, additional files (the basis files) are needed. These files contain the basic response of the detectors, used as reference for comparing the sampled signals. These basic responses have been obtained by scanning the segmented Ge crystal and by Monte Carlo simulation techniques **[25,26,27]**.

The off-line PSA and γ-ray tracking reprocessing is performed on the Grid by replaying the recorded experimental raw data using the adapted AGATA reprocessing software. This software reads all of the data files corresponding to one dataset, process them, and produces the FPD, as shown in Fig. 5.3.

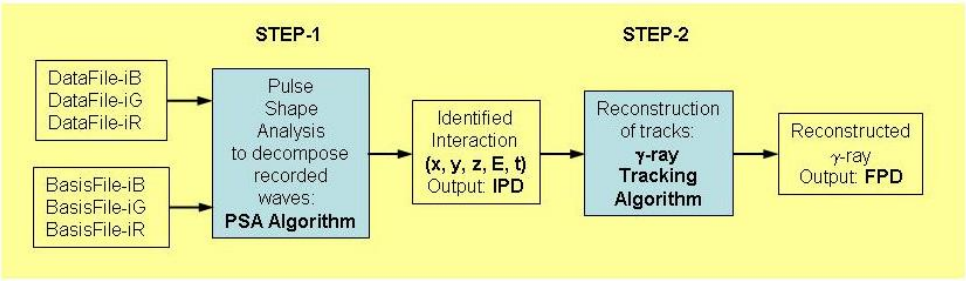

Figure 5.3 : Diagram showing the workflow of the AGATA reprocessing software.

The signals coming from different crystals and corresponding to the same event are identified by their global timestamp, provided by the Global Trigger and Synchronization (GTS) AGATA system.

The reprocessing is performed in two steps, PSA processing and then γ-ray Tracking processing. In case of complexity of one step with regard to the other, the PSA processing and the γ-ray Tracking processing would be performed separately. First, the only PSA is processed producing IPD files, and then the γ-ray Tracking is processed to produce the FPD.

5.3. Running the AGATA Data Processing Software on the Grid

Another important point related with the optimal use of the Grid computing resources concerns the way to deal with processing data through job submission to the Grid.

As shown in Fig. 5.4, the job is prepared at the local User Interface computer. It is assumed here that the AGATA reprocessing software is yet installed on the Grid. The configuration files (if any) are submitted to the Grid, together with the JDL file and the script file. Once the job arrives to a Computing Element and starts running on a given Worker Node, the script file is launched and the AGATA reprocessing software starts running. At this point, the AGATA software interacts with the SE to deal with the input data files. Here, two options are possible.

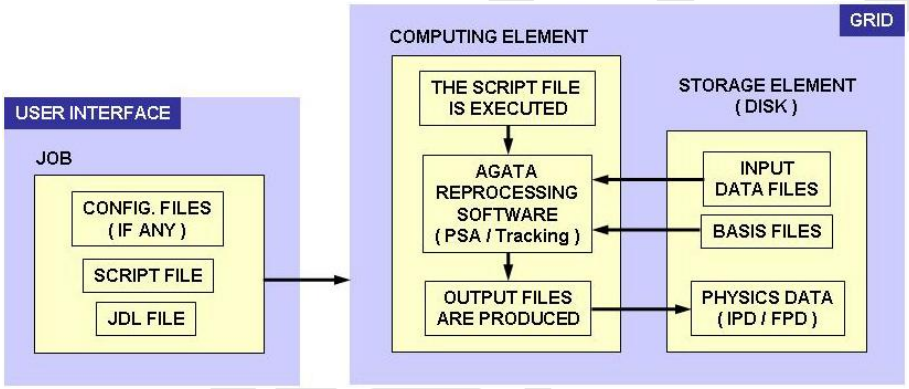

Figure 5.4 : Workflow diagram for a PSA and γ-ray tracking job running on the Grid

The first one is that, before the execution of the AGATA software, the script file initiates previously a copy of the input data files into the WN where the job is running. So, the software can access locally the data and not need any special Grid access protocol. In this option, a data transfer is needed from the SE to the WN, and this query must be included previously in the script file by the user.

The second option needs to deal with data access protocols and GFAL library and allows remote access to input data. In this option, no data transfer is needed. However, the software code has to be modified to include the necessary APIs to use GFAL. It is worth noting that the remote access to data works wherever the job is running. Particularly, the job can be submitted to a CE located at the same Grid site where is located the SE hosting the needed data.

Important note: In both the above options, the performance, in terms of execution time of the AGATA software, is better when both of the SE (where input data are stored) and the CE (where the jobs will run) are located at the same Grid site.

5.4. Example of Running PSA and γ-ray Tracking on the Grid

In the following, an example of running PSA and γ-ray Tracking on the Grid using the Narval emulator is described.

The raw data used for running off-line the PSA and γ-ray tracking analysis on the Grid were obtained from one of the first commissioning experiments performed at INFN-LNL, Legnaro (Italy). The reaction was performed with a ^{30}Si beam with 70 MeV energy, delivered by the tandem of LNL, impinging in a ^{12}C target with a thickness of 0.2 mg/cm^2. One triple cluster detector module of AGATA, namely ATC1, placed perpendicularly oriented with respect to the beam, was used to detect the γ-rays emitted by the recoiling excited nuclei produced in the reaction. A total amount of 13 TB raw data, corresponding to around 10^9 events, was collected and stored on tape at the Tier1 INFN-CNAF site at Bologna (Italy) **[28]**.

The data consists of 36 datasets. Each dataset contains the data files corresponding to the 3 segmented crystals of the triple cluster (Ge:Red, Ge:Green, Ge:Blue). A total of more than 300 data files, each with a size of 14.3 GB, have been recorded for each crystal.

The tests of running the off-line PSA and γ-ray tracking on the Grid have been performed at IFIC-Valencia (Spain) using the gLite middleware **[29]**. The original dataset files are stored on tape at the LCG-Tier1 INFN-CNAF (Castor-CNAF) at Bologna (Italy). Samples of these data have been transferred to the Storages Elements (SE) located at IFIC for further tests. 0.6 TB of data have been copied to tape (Castor-IFIC) and 2.1 TB have been replicated to disk (Lustre-IFIC). The transfers have been done with rates relatively constants ranging between 10 and 12 MB/s.

Mainly, the Grid computing resources from Grid-CSIC **[30]** at IFIC have been used. They consist of sufficient disk storage capacity and a cluster of 50 (8-CPU-cores) computers of 2.8 GHz CPU cycle and 16 GB RAM (2 GB per core) running Scientific Linux 5. Disk space dedicated to the Grid at IFIC is managed by a distributed storage file system, namely Lustre **[31,32]**. A tape storage managed by Castor **[19]**, is also provided at IFIC. Both disk and tape storages are managed by the SRMv2 interface **[20]**.

The first part of the work consisted in adapting the PSA and γ-ray tracking algorithms to run on the Grid, as described in the section 5.2 above. Once this has been done and validated various PSA and γ-ray tracking Grid tasks have been run. A Grid task is defined as a set of PSA and γ-ray tracking jobs submitted to the Grid. Each job reads data from three input data files (14.3 GB size each), and from three basis files (about 350 MB size each).

A first set of tests consisted in running Grid jobs that initiate previously the transfer of a copy of the required input data files from the SE that hosts the data to the Worker Node (WN) where the job run. This is the case for tasks 1 and 2. Task 1 run 14 jobs at the IFIC cluster on 0.6 TB data stored at Castor-IFIC. Task 2 run 46 jobs at the FZK cluster **[33]** on 2.0 TB data stored at INFN-CNAF **[28]**. A second set of tests consisted in running Grid jobs that access directly the data stored at Lustre-IFIC SE. This is the case for tasks 3 and 4, which run 14 and 49 jobs at the IFIC cluster, respectively.

Fig. 5.5 shows the evolution with time of the number of completed Grid jobs for the respective similar tasks 1 and 3 in one hand (left), and, tasks 2 and 4, in the other hand (right). In the same figure is also reported the estimated results obtained for running sequentially, by bunches of 8 parallel jobs, 14 jobs (left) and 49 jobs (right) on a 8-cores computer identical to those used in the IFIC Grid-CSIC cluster, and using all its 8 CPU-cores. These estimated results are obtained by extrapolating the measured execution time of 8 PSA and γ-ray tracking jobs, running in parallel on the 8-cores, to the total number of jobs considered.

Fig. 5.5 indicates that running off-line a task of PSA and γ-ray tracking on the Grid is much more efficient in terms of execution time than running sequentially on a desktop computer, even with 8-CPU-cores fully used, particularly when high amount of data is processed. As shown in Fig. 5.5 (right) running 49 PSA and γ-ray

tracking jobs (on 2 TB data), by bunches of 8 parallel jobs, on a 8-core desktop machine needs around six times more time to complete than running them on the Grid.

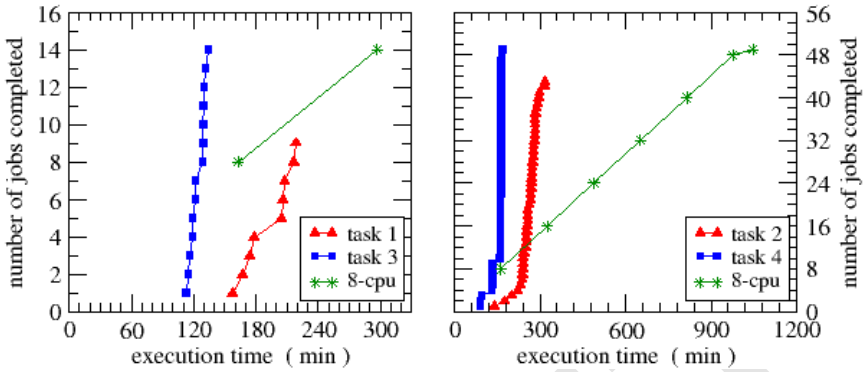

Figure 5.5 : Comparison between the executions of similar PSA and γ-ray tracking Grid tasks, in two approaches: input data are copied to the WN (triangles) and data are accessed directly (squares). Running the tasks on a 8-core computer, by bunches of 8 parallel jobs, is also shown (stars).

Fig. 5.5 shows also that the tasks in which jobs accessed directly the data under the Lustre file system (tasks 3 and 4) have run, respectively, around twice faster than the ones (tasks 1 and 2) that needed a data transfer from the SE to the WN before running. Moreover, tasks 3 and 4 have run with a maximum efficiency (100%), avoiding the failures due to data transfer problems.

These tests have demonstrated that the processing of PSA and γ-ray tracking for AGATA can be run off-line on the Grid with good performances. The best efficiency and execution time are observed for running PSA and γ-ray tracking tasks on high amount of data, with direct access to them using the Lustre file system. In these conditions, and using Grid resources similar to those of IFIC-Grid-CSIC, the execution time of PSA and γ-ray tracking Grid tasks is lesser than 3 hours for processing more than 2 TB data.

Transferring data from a SE to a WN before running, gives also good results, compared to the non-Grid processing. In this case, it is recommended to deal with data files of relatively small size (between 2 and 5 GB) in order to reduce the job failures due to data transfer errors, improving then the efficiency of the executed Grid task. However, in this case, a better performance is reached when the SE (data) and the CE (jobs) are located at the same Grid site.

An estimation based on this work indicates that running the current PSA and γ-ray tracking off-line on the Grid, on up to few tens of TB raw data (which is expected from a standard AGATA experiment), would produce the corresponding reduced data within 1-2 days. Moreover, running PSA and γ-ray tracking off-line allows reprocessing the raw data at any time in case of new processing conditions (new or updated algorithm or version for example), without loss of the original information.

5.5. Example of Running Data Analysis on the Grid

The information containing the reconstructed γ-rays obtained after the PSA and γ-ray tracking reprocessing are saved in the FPD. These FPD are the starting point for the AGATA data analysis performed by the members of an experiment.

The data analysis using FPD may be performed directly on the Grid as the FPD are initially already stored on the Grid and access to them is provided to members of appropriate experiment. In this case, the data analysis

software has to be adapted to run on the Grid and the additional JDL and script files have to be provided by the user. Once this is done, and the software tested and its operation on the Grid validated, the user can submit its jobs to perform his analysis.

In the following an example of performing such data analysis on the Grid is discussed, using the data of an old and real experiment performed with the GASP+ISIS detector system. These data may be taken as equivalent to the AGATA FPD because will be used in the same way.

The data are obtained in a gamma-particle experiment using the GASP+ISIS Ge-detectors arrays. GASP consists of 40 large Ge detectors assembled as a ball covering 4pi space inside which the particle detector ISIS is located around the target. The ISIS detector consists of 40 ΔE-E Si detectors used for light particle detection. The ISIS Si detectors and the GASP Ge detectors are located at the same angles. The experiment consisted in bombarding a target of ^{190}Pt with a beam of ^{12}C delivered by the tandem ALPI of Legnaro on a target made by 2x350 μg/cm^2 thickness. About 80 Gigabytes data, corresponding to around 10^9 events, were recorded on tapes for analysis.

For this work, the data have been copied to the storage Element CASTOR at the IFIC farm. The software used for the analysis performs an event by event sort including required conditions on γ-ray multiplicity and energy as well as particle of interest selection, and produces spectra corresponding at each detection angle and specific γ–γ matrices. Each defined Grid job will deal with the analysis of the data collected in a single run of the experiment and produces the desired output files (spectra and matrices). The output files are merged as soon as they are retrieved from the Grid in such a way that partial results can be looked at before all the experiment is processed.

From the point of view of the execution time, the whole analysis last about 40 minutes using 15 CPU-cores of the IFIC Grid Farm. For comparison, the same data sorting run on a single computer with data stored on a local disk took a bit less time (35 minutes). The fact with AGATA is that the amount of data to analyse (FPD) is expected to be at least one order of magnitude higher than that used for this test (around 250 GB of FPD for 1 TB raw data), due to the high sensitivity and peak to total of AGATA. In this case, the Grid may also be a good tool for data analysis, particularly for experiment with high multiplicity.

5.6. Grid Tools for AGATA users

The submission to the Grid of large number of jobs needs writing scripts to make automatic the management of the job submission and the retrieving of the output files, if any. The model described in section 2.4. is not enough user-friendly for the VO **vo.agata.org**. There is a usability restriction because the user has to deal with JDL syntax. Other issue is the command line job management (steps 10,12) that can discourage an efficient use of Grid resources. For these reasons, it is necessary to use scripts or applications to simplify the job management of the user.

High-level applications, user oriented, providing automatic job management tasks in the Grid, have been developed in the e-Science Group of IFIC-Valencia (Spain). The applications provide to the users a user friendly graphical interface that allow him/her to interact with the Grid without dealing with its complexity. The applications have been adapted to the AGATA requirements and have been used for job submission in the frame of the AGATA data processing and data transfer tests. In case of data transfer, the application is providing full consistency integration with the deployed and tested LFC catalogue (see section 2.2.).

6. THE AGATA GRID RESOURCE REQUIREMENTS

In the following section, the computing resources required for the AGATA project will be estimated, in the basis of the operation of the demonstrator during the commissioning and the first experiments performed at INFN-LNL, and, using the first results of the data reprocessing tests performed using Grid resources **[22]**.

6.1. The Storage Capacity Requirements for AGATA

During its construction at INFN-LNL at Legnaro (Italy), the demonstrator has been used, with various triple clusters mounted, to perform the primary commissioning experiments. During the year 2009, seven experiments have been performed with one or two triple clusters. A total of around 40 TB of data have been collected for the commissioning experiments of the year 2009 and more than 70 TB have been collected for the seven experiments performed during the first semester of 2010. The amount of data collected with the demonstrator, using four triple clusters, vary between 3 and 30 TB, depending of the kind of experiment performed. One can note that the data collected could differ of one order of magnitude when going from experiments with low multiplicity (nucleon transfer reaction) to experiments with high multiplicity (high spin population).

Tables 6.1a and 6.1b show the data collected in different experiments **[3]**:

Exp.	Week-12	Week-22	Week-27	Week-43	Week-46	Week-49	Week-06*
Size	15 TB	300 MB	2.2 TB	14 TB	1.5 TB	6.7 TB	1.9 GB

Table 6.1a : Commissioning data

Exp.	Week-07	Week-09	Week-19	Week-21	Week-24	Week-25	Week-28
Size	3.4 TB	2.9 TB	8.9 TB	20 TB	4.5 TB	3.4 TB	30 TB

Table 6.1b : Experimental data

Note: In the following it is assumed that the maximum amount of data collected in one of the experiments can reach 30 TB.

According to the model, when the first experiment is going on, the disk storage of the DAQ farm is receiving raw data and is being continuously filled. The disk storage capacity of the DAQ farm have to cope with the highest amount of data collected in a high multiplicity experiment, around 30 TB collected with the demonstrator. The raw data transfer to the Tier1 tape storage must start as soon as collected data are available (some runs complete). The transfer of the data of an experiment is being performed while new collected raw data are being recorded on the DAQ farm storage. Once the first experiment completes, the next one may start straightforward, before the entire data transfer of the first experiment finished. The minimum capacity of the DAQ farm must then be enough (around 60 TB) to cover the data storage of two successive high multiplicity experiments. This is the storage capacity that allows the DAQ farm operating normally for successive experiments. In normal operation, when the third experiment starts, the data of the first experiment must have been already deleted from the DAQ farm disk because completely transferred and saved on tape at the Tier1. This way ensures that enough disk space is available for two most consuming disk space successive experiments.

However, for safe operation, the disk storage of the DAQ farm may be increased to be able to store the data of an additional experiment in case of temporary failure of data transfer to the Tier1.

Then, from the above, the expected disk storage for the DAQ farm is estimated to be at least 60 TB, 90 TB for safe operation.

The disk storage at the Tier1 and Tier2 sites must be used efficiently in order to optimize its cost. For that, all of the raw data of an experiment (particularly experiments with high multiplicity) is copied from tape to disk storage for reprocessing only when the processing PSA+Tracking programs are tested and ready to run. These programs are previously tested and finalized using only a small sample of the raw data (1 or 2 TB) copied to disk.

If an average time of up to 4 months is estimated to be required for performing this pre-reprocessing, the data expected to be stored at the same time on the Grid disk for pre-reprocessing is equivalent to 16 experiments (if successive experiments are performed at a rate of one experiment per week). If each experiment needs around 2 TB data for these pre-reprocessing, the total disk storage required is then of 32 TB during 4 months.

The first of these 16 experiments is expected to complete the data pre-reprocessing within the first 4 months and then starts the final reprocessing of the whole data of the experiment, which, in the worst case, needs 30 TB disk space. This final reprocessing is estimated to complete within few days and up to one week. But, two weeks are considered for a large estimate (any unexpected problem can delay the reprocessing). With this assumption, 2 to 3 experiments would need to be reprocessed at the same time. Then, 60 to 90 TB additional disk space would be needed.

One has to keep in mind that this disk space is for the whole AGATA experiments storage and then can be distributed across various Grid sites. The point here is that AGATA must provide at least 3 Grid sites in order to face at least 3 parallel reprocessing of high multiplicity experiments. Concentrating all of the computing resources in one Grid site is to be avoided because creates inefficiency and lower the performance.

Once the data reprocessing completes the raw data are deleted and the disk space is get back for the reprocessing of the following experiments. This process of fill-delete of raw data on/from the disk space allows to use the same disk space to reprocess the raw data of all of the experiments. From the above described experimental conditions and operation, an estimate of the disk space requirement is around 120 TB, reusable for the data reprocessing.

Note that as the disk storage at the DAQ farm and at the Grid sites are non permanent storage for raw data, the same disk is reusable and no more capacity is needed during many time of the life of AGATA (few years), assuming that the maximum data collected for an experiment is 30 TB. The correction should be done when using the full AGATA, if more than 30 TB are collected in a single experiment. In this case more disk space should installed at that moment.

The FPD produced by the reprocessing for analysis will be stored permanently on Grid. Consequently, the disk storage dedicated to these FPD will increase with time as more experiments are performed. The FPD are estimated to occupy from 1/5 to 1/20 of the raw data after reprocessing.

For the FPD, it is difficult to estimate the required disk capacity necessary for their storage because of the following reasons:

- For low multiplicity experiments, up to 4-5 TB raw data, the produced FPD will occupy few hundreds of GB space and then the members of the experiment will probably prefer to download the data to their home institute for analysis. In this case, only one copy of the FPD may be kept on the Grid.
- For high multiplicity experiments, above the 5 TB raw data, the produced FPD will occupy few TB of disk space, and it is then more efficient to perform their analysis directly on the Grid. In this case, at least two copies of the FPD are kept on the Grid for redundancy to secure the data.
- During the next years, for the next phases of AGATA, the type of experiments to be performed is not yet well established. The expected campaigns of experiments could deal with low multiplicity experiments performed with radioactive beams or with high multiplicity experiments performed with stable beams, or both of them. As we have observed with Demonstrator, one has to remember that the amount of data collected in a high multiplicity experiment may be of one order of magnitude higher than that collected in low multiplicity experiment. This would affect in the same way the disk storage capacity to deal with this.

However, a rough attempt is done using the sample of data production already collected by AGATA in the 7 experiments performed during the first semester of 2010. These data (table b) show that 14% of the experiments produced 30 TB, 14% produced 20 TB, 15% produced 10 TB, and 57% produced 4 TB. Applying these ratios to around 30 experiments performed a year in normal operation of AGATA, the estimated disk space

required to store the FPD of the whole year is around 125 TB, taking into account that for all the experiments two copies of the FPD are kept on the Grid for redundancy. Then disk space may be added at the time it is needed.

In summary, the total disk space on the Grid required for the normal operation of AGATA during one year, is estimated to around 250 TB, in the conditions of operation described above, i.e. 30 successive experiments with a maximum of 30 TB data in 14% of the experiments. These numbers for disk capacity requirements assume that 100% of the disk space is used. In fact an additional disk capacity (10-15%) has to be taken into account to make the operation flowing well.

The data produced by the end-user physicists as an output of their analysis on the Grid using the FPDs are not taken into account because they require small amounts of space. Moreover these data will probably be downloaded to local computers for analysis right after their production. Then, depending of the number of active users, a small capacity of the Grid disk space (around 0.3 TB per user), with the write permissions, may be provided to the users at each Grid site.

6.2. The Computing Power Requirements for AGATA

The main areas of computing activity that have to be considered are the following :

- The PSA processing, repeated few times, to produce IPD, followed by the γ-ray Tracking processing, which may be repeated many times for each experiment in order the reach the best precision for the FPD. This reprocessing (PSA+Tracking) is performed with a small sample of raw data.
- When the reprocessing methods and parameters are tested and finalized for a given experiment, run the reprocessing with the entire raw data of the experiment and produce the FPD.
- The data analysis run on the Grid using the produced FPD. The data analysis is more uncertain in its requirements because it is chaotic. The user may want to run its analysis programs on the Grid. It is assumed that at least the analysis of experiments with high amount of data will be run on the Grid (assuming around 30% of the experiments). Here, the data analysis at the home institute is not taken into account.

The tests of running PSA and g-ray Tracking on the Grid using the Narval emulator with the commissioning data obtained with one triple cluster have shown a very encouraging results **[23]**. These tests have been performed using sufficient disk storage capacity and a cluster of 8-CPU cores computers of 2.8 GHz CPU cycle and 16 GB RAM (2 GB per core) running Scientific Linux 5. The PSA+Tracking (Narval emulator) processing 2 TB of data and running on the Grid using 50 cores lasted 3 hours. One has to note that by using Lustre during theses tests, direct data access has been use. Considering that the result shown in Fig. 5.5 (right-task-4) has been obtained with the best conditions, let us assume that one need 6 hours to process 2 TB in normal conditions.

Assuming a linear extrapolation, in the same conditions as described above, 30 TB raw data would be reprocessed in 90 hours (about 4 days), and in around 45 hours if the number of CPU cores is 100. These are reasonable times to reprocess such a huge amount of data. However, it should be taken into account that, at normal operation of AGATA, raw data of probably up to 3 experiments would be running at the same time across the Grid. To avoid overloading a particular site, raw data of consecutive experiments must be distributed over different Tier1 and Tier2 sites. At least 3 Grid sites must be provided by AGATA. In this case, about 50-60 cores per site (6-8 computers of 8-cores each) may reprocess the data of a quite big experiment, in a reasonable time (3-4days), and in shorter time if more CPU cores are used.

Moreover, it should also be taken into account that probably some large amount of FPD would be analyzed on the Grid. In this case, additional resources must be provided by the Tiers sites. The required CPU resources for data analysis on the Grid is estimated using the results obtained by running a real GASP data analysis on the Grid. 80 GB of data collected in an experiment using the GASP array spectrometer coupled to the particle detector ISIS have been processed in 40 minutes on the Grid to produce spectra and matrices. This processing used 15 CPU-cores. Then, an average estimate assuming similar time processing of both the GASP+ISIS events

and the FPD, it is expected that around 1 TB of FPD would be run in about 8 hours in the same conditions and 6 TB in around 2 days. One day if the number of CPU-cores is 30. An additional 30 CPU-cores would make an AGATA Grid site operating in a good way with reprocessing and data analysis.

It is important to note that although these estimates of CPU are based on the maximum data collected in one "big" experiment, no particular additional CPU is needed to cover the rest of the experiments, taking into account the following:

- The CPU occupancy to reprocess 30 TB of raw data will be used for only few time (around one week), after which the CPU is again available.
- The time last to perform a data analysis on FPDs is expected to be quite small comparing with the reprocessing time, in such a way that CPU is quickly available for an analysis after have been used from a previous one.
- The CPU resources have to be used optimally to improve the performances and efficiency.

Then, an AGATA Grid site with 80-100 CPU-cores will cover the needs in computing power for data reprocessing and data analysis in normal operation, within the conditions described above. However, at least 3 Grid site with that size have to be provided by the AGATA project.

6.3. Ramp-up and Resource Requirements Evolution

From a cost point of view, the best way to purchase the required resources would be 'just in time'. When the resources are really needed, but taking into account the type of experiment to be performed particularly for the disk storage requirements, but also processing. There is therefore a requirement for early installation of both CPU and disk capacity.

As a summary, the previous attempt of Grid computing resources requirements needed by AGATA are made assuming the following:

- AGATA phase one (Demonstrator) is used.
- 30 successive experiments are performed in a year (one experiment per week).
- 14% of the experiments produce 30 TB, 14% produce 20 TB, 15% produce 10 TB, 57% produce 4 TB.
- Maximum amount of raw data produced in one single experiment is 30 TB.
- Data reprocessing (PSA+Tracking) is performed first with a small raw data sample (2 TB) then run on all of the raw data corresponding to the experiment (30 TB max). Afterwards, the raw data are deleted.
- Two copies of the produced FPD are kept permanently on the Grid.

In these conditions, among the Grid resource requirements estimated (disk capacity and computing power), only the disk storage required for the permanent storage of the FPD has to be incremented yearly, as a function of the number and type of experiments to be performed. The disk space used by the raw data and the CPU are reusable. In addition, even the disk space dedicated to the FPD may be managed in such a way that old experiments (of two or three years old) may be deleted or saved by copying it to tape in order to recover disk space.

For the next two years, AGATA may operate with the Grid computing resources estimated in this document if the experiments expected to be run at GSI are of low multiplicity, even with more crystals. Operating 15 triple clusters leads to a raw data collection of about 16 TB data comparing to the 4 TB obtained with Demonstartor. In case that high multiplicity experiments are planned, additional computing resources have to be provided at a time.

6.4. The Networking Requirements

The traffic is based on the planned flow of raw data from the Data Production site (experiment) to the Tier1 sites, between the Tier1 sites themselves, and from the Tier1s to the Tier2s. It is also based on the planned flow

of the FPD between all of the Tier1 and Tier2 sites. The precise bandwidth required depends on the network topology assumed to link the Data Production site to the Tier1s and the Tier1s to Tier2, but also the inter-Tier1s and inter-Tier2s sites.

The throughput for the data transfer from the Data Production site to the Tier1 sites must be at least 100 MB/s. This value of throughput allows performing the transfer of all of the raw data corresponding to the maximum data collected in an experiment with high multiplicity (around 30 TB) within one week, which is the average duration of an AGATA experiment. The throughput is defined here as the volume of data transferred in a given time period.

The bandwidth required between the Tier1s and the Tier2s will depend on the sites concerned, but in terms of the files to be available for general use at each site, it will be typically 10 MB/s or less. The traffic associated with user jobs is not included, but may be important, depending on the activity on the Grid of these users.

During the running PSA+Tracking tests on the Grid (see section 5.4), throughputs have been measured between the different sites where jobs have run. Values ranging between 3 MB/s and 9 MB/s have been observed for transfers CNAF → FZK, CNAF → LYON and IFIC → Manchester. However, it is worth noting that the throughputs may fluctuate quite a lot, and, relatively much higher values have been observed during these tests. Values of up to 30 MB/s have been measured for CNAF → FZK-1 data transfers, 39 MB/s for CNAF → LNL.

Recently (august 2010), values of throughputs up to 23 MB/s have been measured for CNAF (SE) → LYON (SE) data transfer using files of 14 GB size. The throughputs measured for the transfer of such files from CNAF (SE) to the various sites (WN) of France has shown values ranging from 2 MB/s to 12 MB/s equally for Lyon, Strasbourg and Paris.

6.5. The Current AGATA Grid Computing Resources

Presently, the Grid sites involved in the AGATA VO **vo.agata.org** consist in two Tier1 sites and four Tier2 sites. The Tier1 sites, INFN-CNAF Bologna in Italy and CC-IN2P3 Lyon in France, are providing only tape storage with a relatively small disk cache space, 10 TB at CC-IN2P3 for example. The Tier2 sites, namely IPHC-Strasbourg, IPN-Lyon, IPN-Orsay and very recently (January 2011) IFIC-Valencia, are providing both SEs (disk space) for data storage and CEs for data processing.

The current (last week of January 2011) Grid computing resources available for the VO **vo.agata.org** are reproduced in the following table 6.5a for the CEs and table 6.5b for the SEs.

Total CPU	Free CPU	Total Jobs	Running Jobs	Waiting Jobs	Computing Element
1504	4	0	0	0	sbgce2.in2p3.fr:8443/cream-pbs-vo.agata.org
848	823	0	0	0	ce03.ific.uv.es:8443/cream-pbs-short
848	823	0	0	0	ce03.ific.uv.es:8443/cream-pbs-agataL
800	198	0	0	0	ipngrid04.in2p3.fr:8443/cream-pbs-sdj
640	38	0	0	0	ipngrid04.in2p3.fr:8443/cream-pbs-agata
800	198	0	0	0	ipnls2001.in2p3.fr:2119/jobmanager-pbs-sdj
640	38	0	0	0	ipnls2001.in2p3.fr:2119/jobmanager-pbs-agata
664	332	0	0	0	lyogrid07.in2p3.fr:8443/cream-pbs-vo.agata.org
664	332	0	0	0	lyogrid02.in2p3.fr:2119/jobmanager-pbs-vo.agata.org
1504	4	0	0	0	sbgce1.in2p3.fr:2119/jobmanager-pbs-vo.agata.org

Table 6.5a : Available Grid queues (CEs) for **vo.agata.org** at January 2011

Available Space (kB)	Used Space (kB)	Type	Storage Element
10737418240	3409924231	n.a	ccsrm02.in2p3.fr
1000000000000	500000000000	n.a	srm-v2.cr.cnaf.infn.it
36511000000	321996000000	n.a	srm-v2.cr.cnaf.infn.it
149687126458	n.a	n.a	storm-fe-archive.cr.cnaf.infn.it
138315780	3024162	n.a	srmv2.ific.uv.es
87419358331	355120627695	n.a	sbgse1.in2p3.fr
4079807186	2277618215	n.a	ipnsedpm.in2p3.fr
10165077934	83553231248	n.a	ipnsedpm.in2p3.fr
4990008918	10796826	n.a	lyogrid06.in2p3.fr

Table 6.5b : Available Grid storage (SEs) for **vo.agata.org** at January 2011

7. THE AGATA GRID USER SUPPORT

The aim of the AGATA Grid User Support is to provide the necessary support to the AGATA collaboration members in their use of the AGATA Grid Infrastructure. This can cover the following point :

- Access the Grid Resources : create/renew Grid certificate, VO membership, use of the AGATA software on the Grid,…
- Data management : use of the LFC catalogue, lcg-xx and lfc-xx command lines,…
- Job management : use of the gLite command lines, glite-wms-xx, file configurations,…
- data reprocessing and data analysis Grid support
- technical Grid support
- software Grid support
- support for the Grid use in general

These tasks have to be organized and the support may be provided through several tools, like web pages, e-mails, tutorials, etc… In case of problems with an AGATA Grid site, the GGUS (Grid Global User Support) **[34]** may be used to post tickets or alternatively use the AGATA grid-support e-mail contacts of the sites.

REFERENCES

[1] AGATA official web site : http://www-win.gsi.de/agata/
[2] INFN-LNL : http://www.lnl.infn.it/, http://agata.lnl.infn.it/
[3] http://csngwinfo.in2p3.fr/mediawiki/index.php/Grid_HowTo
[4] LCG : http://lcg.web.cern.sh/lcg
[5] I. Foster and C. Kesselman, "The GRID Blueprint for a New Computing Infrastructure", Morgan Kaufmann Publishers, Inc., San Francisco, USA, ISBN 1-55860-475-8
[6] VO vo.agata.org : cclcgvomsli01.in2p3.fr:8443/voms/vo.agata.org
[7] EGI-InSPÎRE Project : http://www.egi.eu/projects/egi-inspire
[8] EGEE Project : http://www.eu-egee.org/
[9] E. Laure et al., "Programming the Grid with gLite", Computational Methods in Science and Technology 12(1), 33-45 (2006)
[10] Globus : http://www.globus.org/ogsa/
[11] WMS : https://edms.cern.ch/document/572489/1
[12] LB : http://egee.cesnet.cz/en/JRA1/LB/documentation.php
[13] BDII : http://twiki.cern.ch/twiki/bin/genpdf/EGEE/InformationSystem
[14] VOMS : http://www.globus.org/grid_software/security/voms.php
[15] LFC : http://glite.web.cern.ch/glite/packages/R3.2/s15_x86_64_/deployment/glite-LFC_mysql/ \ glite-LFC_mysql.asp
[16] GLUE schema : http://www.ogf.org/documents/GFD.147.pdf
[17] FTS : https://wiki.chipp.ch/twiki/pub/LCGTier2/FTSlinks/transfer.pdf
[18] RFT : http://dev.globus.org/wiki/Reliable_File_Transfer
[19] Castor : http://cerncourier.com/cws/article/cern/31529
[20] StoRM : http://storm.forge.cnaf.infn.it
[21] JDL : http://www.grid.org.tr/servisler/dokumanlar/DataGrid-JDL-HowTo.pdf
[22] M. Kaci et al., 4th Iberian Grid Infrastructure Conference Proceedings, Braga, Portugal, may 24-27, 2010
[23] V.Méndez et al., "A Decentralized Deployment Strategy and Performance Evaluation of LCG File Catalogue Service" To be publicated in Journal of Grid Computing
[24] ClassAd : http://www.cs.wisc.edu/condor/classad
[25] E. Farnea et al., Submitted to Nucl. Inst. and Meth. A
[26] P. Medina et al., "A simple method for the characterization of a HPGe detector" in Proc. IMTC - 2004
[27] L.Nelson et al., Nucl. Inst. and Meth. A 573 (2007) 153
[28] INFN-T1, Bologna, Italy : http://grid-it.cnaf.infn.it/
[29] The gLite project : http://glite.web.cern.ch/glite/
[30] The Grid-CSIC web site : http://www.grid.csic.es
[31] The Lustre file system : http://www.lustre.org
[32] Lustre at IFIC : https://twiki.ific.uv.es/twiki/bin/view/Atlas/LustreStoRM
[33] FZK-LCG2, FZK, Karlsruhe, Germany: http://www.gridka.de
[34] GGUS : http://www.ggus.org