GRID COMPUTING APPLIED TO OFF-LINE AGATA DATA PROCESSING

M. KACI

mohammed.kaci@ific.uv.es

2nd EGAN School, 03-07 December 2012, GSI Darmstadt, Germany

GRID COMPUTING TECHNOLOGY

THE EUROPEAN GRID: HISTORY

In 2001 the project named "Research and Technological Development for an International Data Grid“ known as the European Data Grid Project was funded for three years.

A major motivation behind the concept was the massive data requirements of the Large Hadron Collider (LHC) project of the European Organization for Nuclear Research (CERN).

On 1 April 2004 the Enabling Grids for E-Science in Europe (EGEE) project was funded by the European Commission , led by the information technology division of CERN. It has been extended by EGEE-II, and followed by the EGEE-III which ended in April 2010

The Worldwide LHC Computing Grid (WLCG) continued to be a major application of EGEE technology.

A middleware software package known as gLite was developed for EGEE.

By 2009 the governance model evolved towards a European Grid Infrastructure (EGI), building upon National Grid Initiatives (NGIs).

WHY GRID COMPUTING

Science has become increasingly based on open collaboration between researchers across the world.

It uses high-capacity computing to model complex systems and to process experimental results.

In the early 21st century, Grid computing became popular for scientific disciplines such as high-energy physics, bioinformatics to share and combine the power of computers and sophisticated, often unique, scientific instruments in a process known as e-Science

Production of high amount of data and computing power needed to process it.

Not feasible to store at a central point.

Distribute resources among participant centers

- Centre puts its computing and storage resources (helps to share costs)
- Data is distributed among centers, always available (replicas)
- Everybody can access remote resources, redundancy services

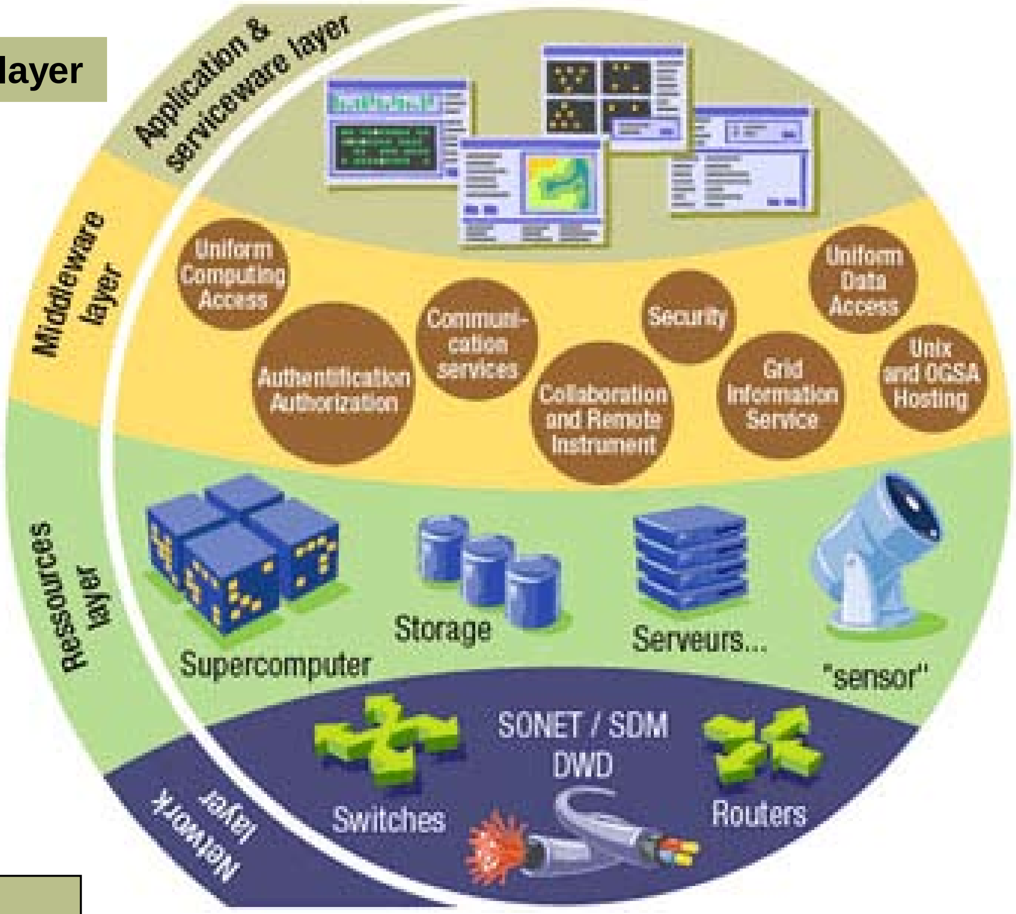
Need technology to access these resources in a coherent manner

- Users belong to a common Organizations (Virtual Organization)
- Secure access and trustworthy relations

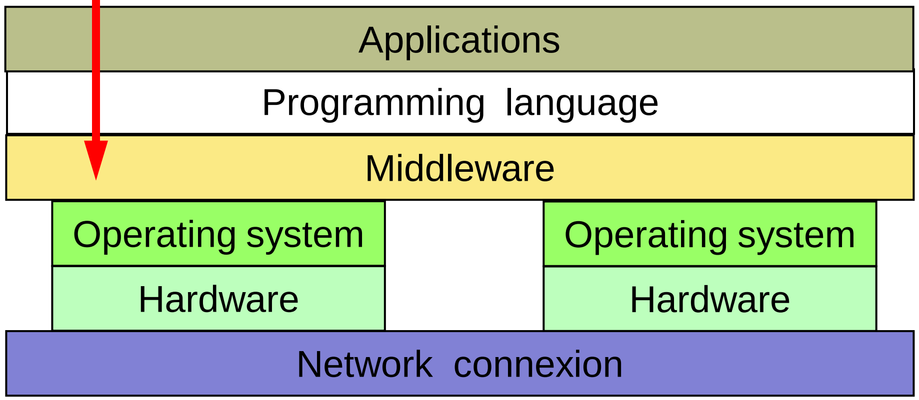
ANATOMY OF THE GRID (I)

Application layer

Middleware layer
sits between Apps and OS to
provide basic access services



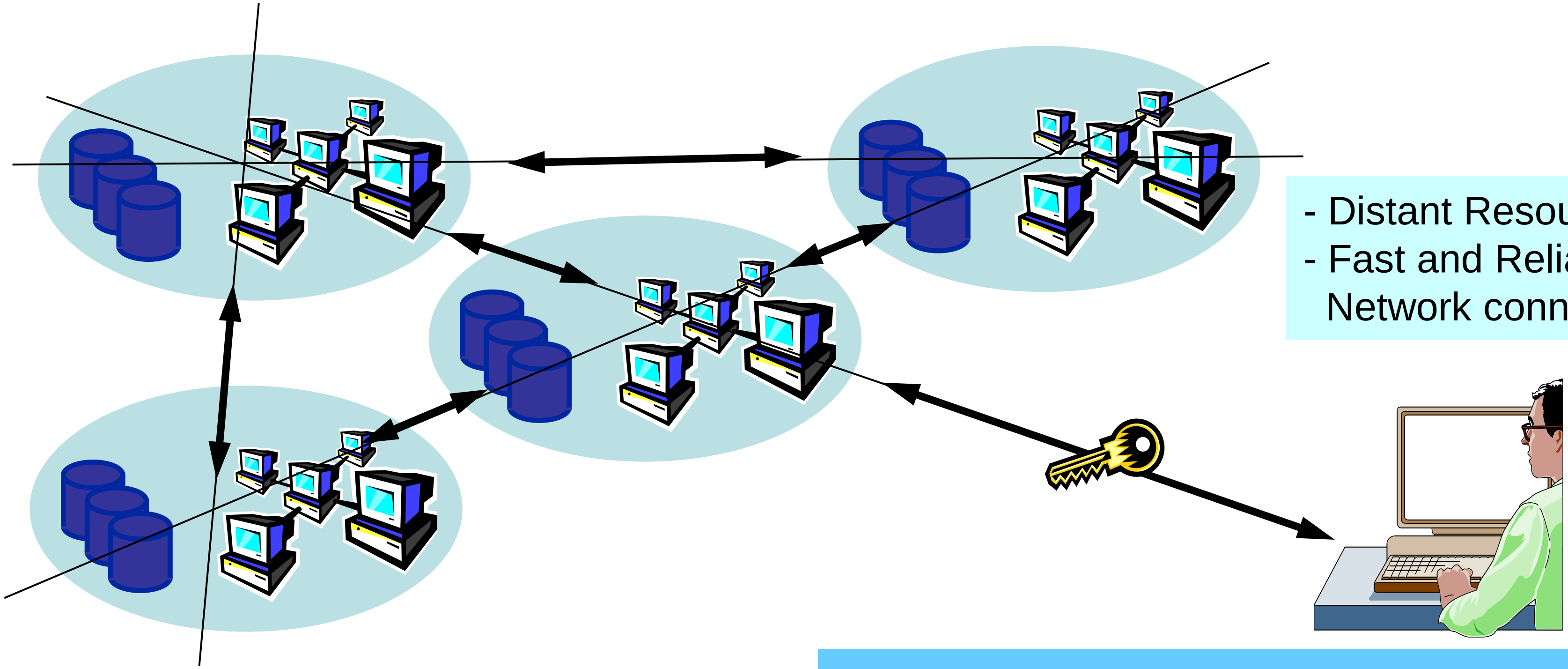
Resources layer
computing, storage, instruments



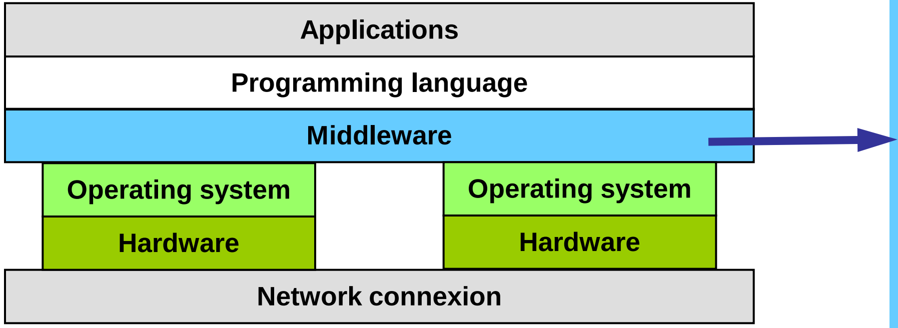
Network layer

ANATOMY OF THE GRID (II)

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).



- Distant Resources
- Fast and Reliable Network connexion



- Find the right site where to run a job
- Optimize the use of the Grid resources
- Organize the eficient access to data
- Authenticate the elements of the Grid
- Execute the tasks
- Monitor the evolution of jobs execution
- Retrieve the results when jobs are done

THE gLite MIDDLEWARE (I)

gLite is a middleware computer software project for the WLCG/EGI Grid.
Now it is part of the European Middleware Initiative (EMI) project.

gLite provides
- a framework for building applications tapping into distributed computing and storage resources across the internet.
- a set of common services to access remote resources in a coherent manner:
  Security Services, User Interface, Computing Element, Storage Element,
  Information Service, and Workload Management

SECURITY: security is based on

Authentication : are you who you claim to be?
Authorization : do you have access to the resource you are connecting to?

The gLite user community is grouped into Virtual Organizations (VOs). A user must join a VO supported by the infrastructure running gLite to be authenticated and authorized to using Grid resources.

To authenticate himself, a user needs to have a valid digital X.509 certificate issued by a Certification Authority (CA) trusted by the infrastructure running the middleware.

The authorization of a user on a specific Grid resource can be done through the Virtual Organization Membership (VOMS)

THE gLite MIDDLEWARE (II)

USER INTERFACE :

The access point to the gLite, and then to the remote computing resources, is the User Interface (UI). This is a machine where users have a personal account and where their user certificate is installed.

From a UI, a user can be authenticated and authorized to use the EGI computing resources.

The user can access the functionalities offered by the Information, Workload and Data management systems.

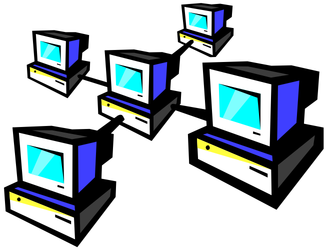
The UI provides the CLI tools to perform some basic Grid operations (practice sessions):

- list all the resources suitable to execute a job
- submit jobs for execution
- cancel jobs
- retrieve the output of finished jobs
- show the status of submitted jobs
- retrieve the logging and bookkeeping information of jobs
- copy, replicate and delete files from the Grid
- retrieve the status of different resources from the Information System

You can also compile your programs and submit the corresponding jobs from the UI.

THE gLite MIDDLEWARE (III)

COMPUTING ELEMENT :



A Computing Element (CE) is some set of computing resources localized at a site (i.e. a cluster, a computing farm).

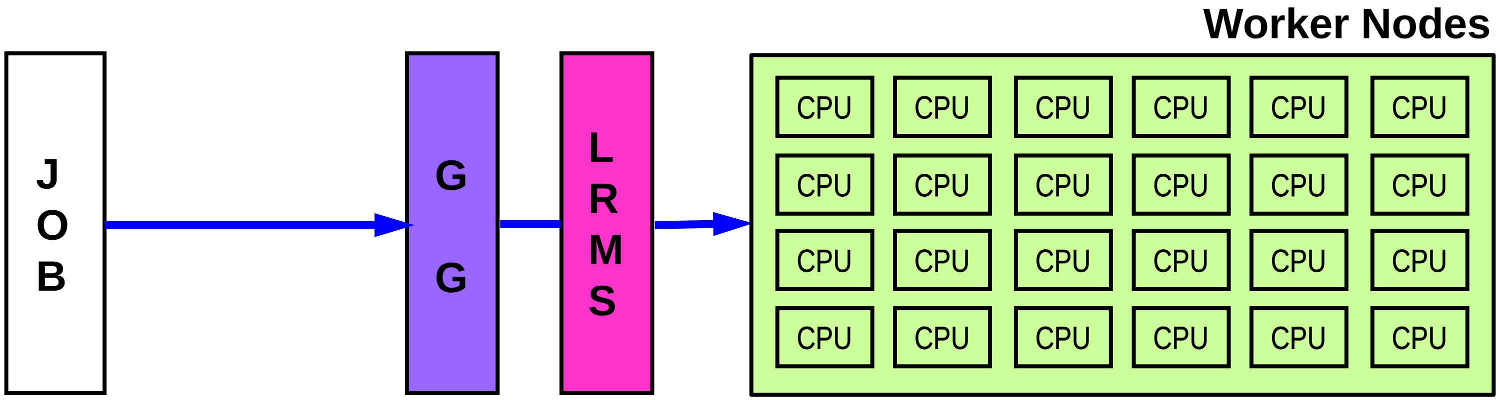
A CE includes :

- a Grid Gate (GG) which acts as a generic interface to the cluster
- a local Resource Management System (LRMS) (sometimes called batch queues system) and the cluster itself
- a collection of Worker Nodes (WNs), the nodes where the jobs are run

The GG is responsible for accepting jobs and dispatching them for execution on the WNs via the LRMS.

A site can have several Ces, grouping homogeneous WNs

Jobs are queued at the batch system until they can be executed



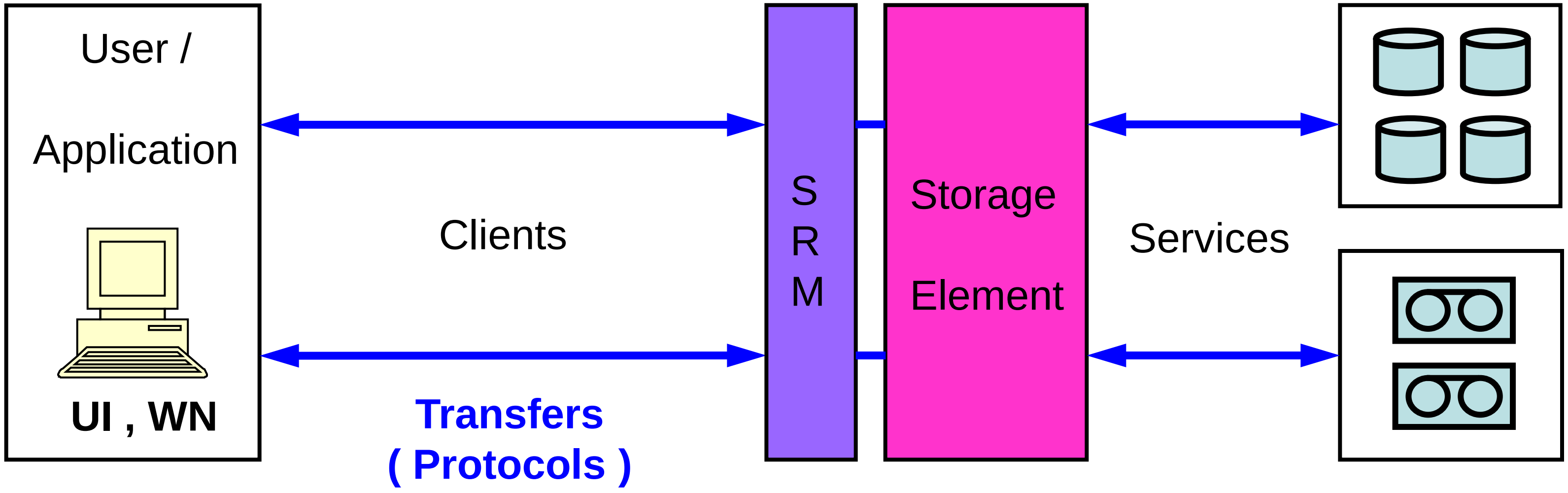
THE gLite MIDDLEWARE (IV)

STORAGE ELEMENT :

A Storage Element (SE) provides uniform access to data storage resources.

The SE may control simple disk servers, large disk arrays or tape-based Mass Storage Systems (MSS).

Most storage resources are managed by a Storage Resource Manager (SRM), a middleware service providing capabilities like transparent file migration from disk to tape…



SRM protocol : used for the Storage Management
GSIFTP protocol : used for the Data Transfers
RFIO, GSIDECAP protocols : used for local or remote data access

THE gLite MIDDLEWARE (VI)

Few Words on Data Management on the Grid :

Data management is about specifically “big files”

- bigger than 20 MB
- In the orden of hundreds of MB
- Optimized for working with this big files

Generally speaking a file in the grid is

- Read only
- Cannot be modified, but
- Can be deleted, so replaced
- Managed by the VO, which is the “owner” of the data
- Means that all members of the VO can read the data.

Grid File Catalogue LFC :

Used to Organize the files on the Grid

THE gLite MIDDLEWARE VII)

INFORMATION SERVICE :

The Information Service (IS) provides information about the WLCG/EGI Grid resources and their status.

This information is essential for the operation of the whole Grid, as it is via the IS that resources are discovered.

The published information is also used for monitoring and accounting purposes

WORKLOAD MANAGEMENT :

The purpose of the Workload Management System (WMS) is to orchestrate the Job management on the Grid. It accepts user jobs, assigns them to the most appropriate CE, records their status and retrieves their output.

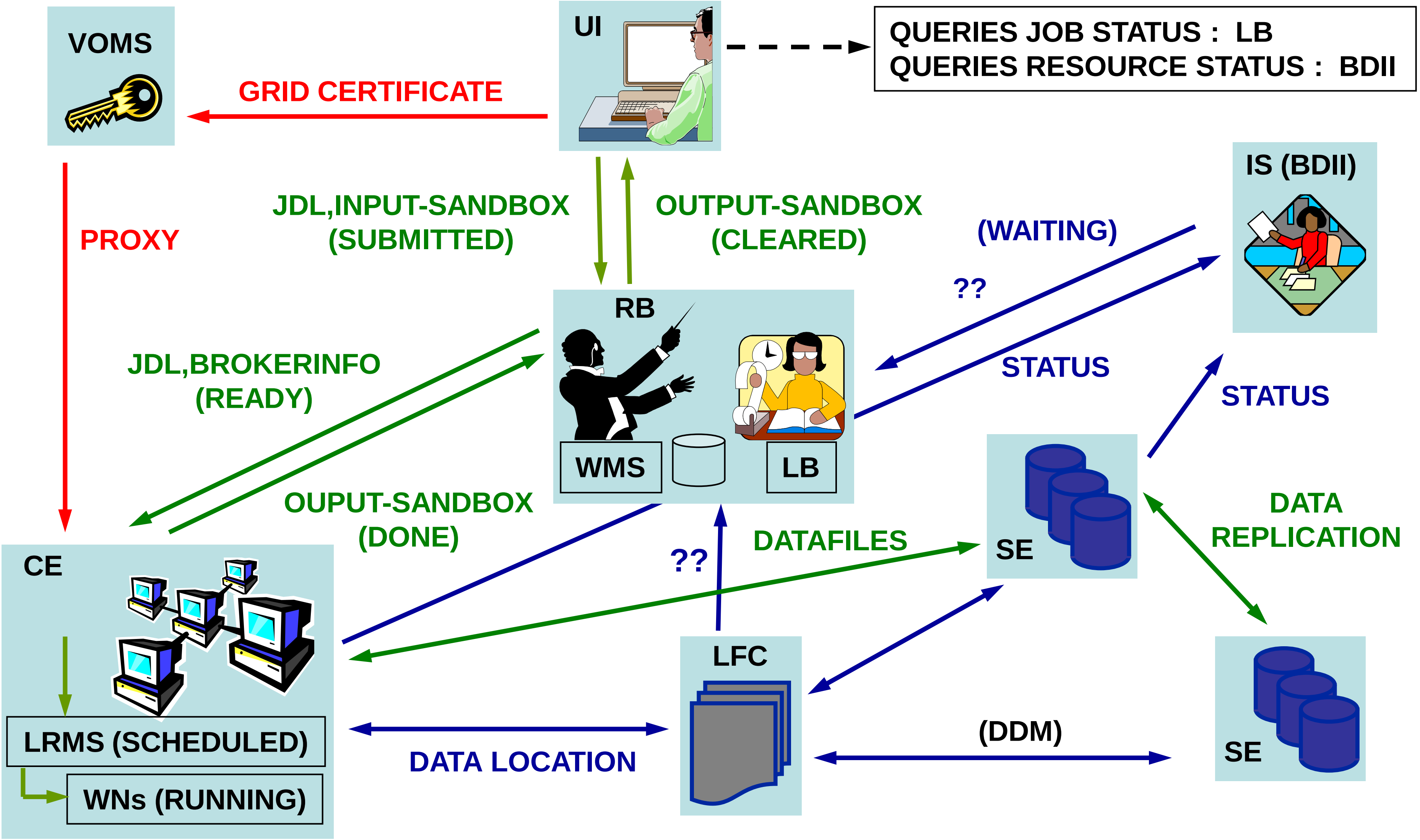
Jobs to be submitted are decribed using the Job description Language (JDL), which specifies, for example,

- which executable to run and its parameters
- files to be moved to and from the WN on which the job is run
- input Grid files needed
- any requirements on the CE and the WN

The logging and bookkeeping service (LB) tracks jobs managed by the WMS. It collects events from many WMS components and records the status and history of the job.

THE gLite MIDDLEWARE (VIII)

Job Flow, Status of the Job



APPLICATION TO
THE OFF-LINE AGATA DATA REPROCESSING

THE ADVANCED GAMMA TRACKING ARRAY PROJECT

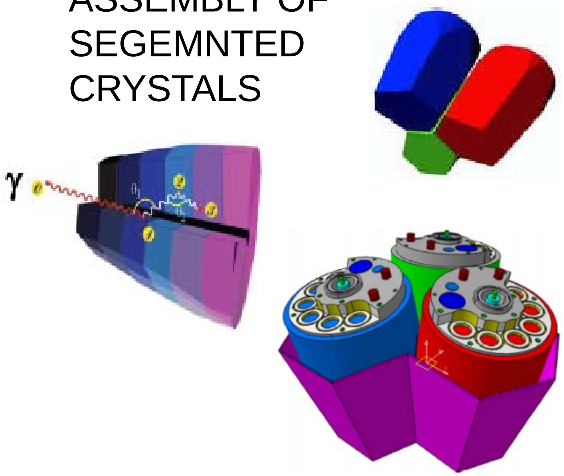
13 European Countries, more than 40 institutions http://www-win.gsi.de/agata/

Build an International experimental facility for Nuclear Physics Research, in particular the exploration of Nuclear Structure at the extremes of isospin, mass, angular momentum, excitation energy, and temperature.

AGATA: a movable Research Instrument to be coupled to stable and exotic ion beam accelerators in European host laboratories:

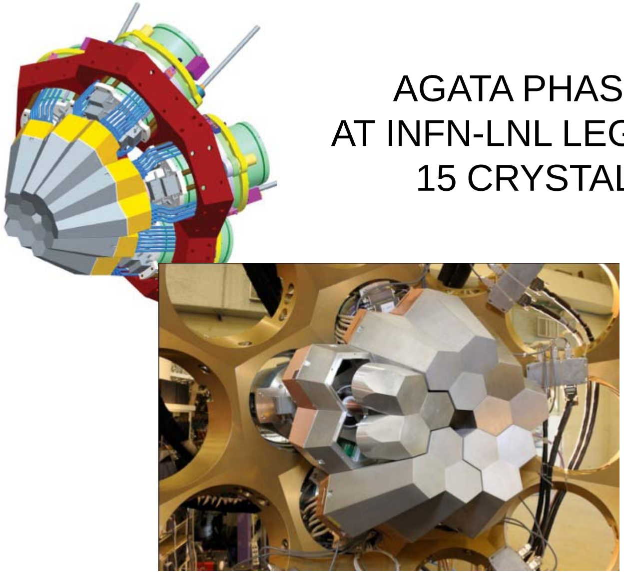
INFN-LNL Legnaro; FAIR/GSI Darmstadt; SPIRAL2/GANIL Caen; ILL Grenoble; REX-ISOLDE/CERN;

ENCAPSULATION AND ASSEMBLY OF SEGEMNTED CRYSTALS



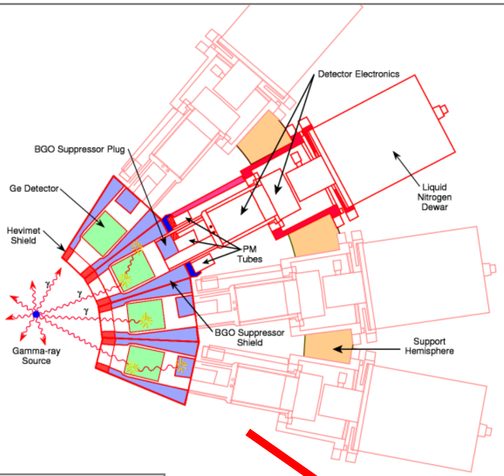
DETECTOR MODULE (CLUSTER)

AGATA PHASE-1
AT INFN-LNL LEGNARO
15 CRYSTALS

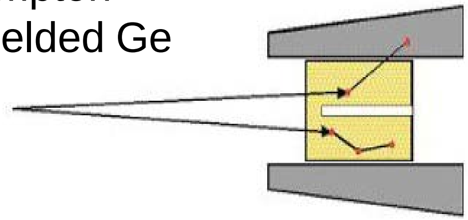


THE GAMMA-RAY SPECTROMETER ARRAYS

Previous Generation: GAMMASPHERE, EUROBALL



Compton
Shielded Ge



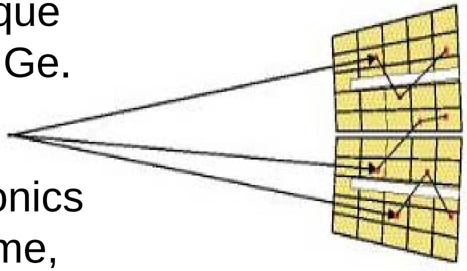
New Generation: AGATA

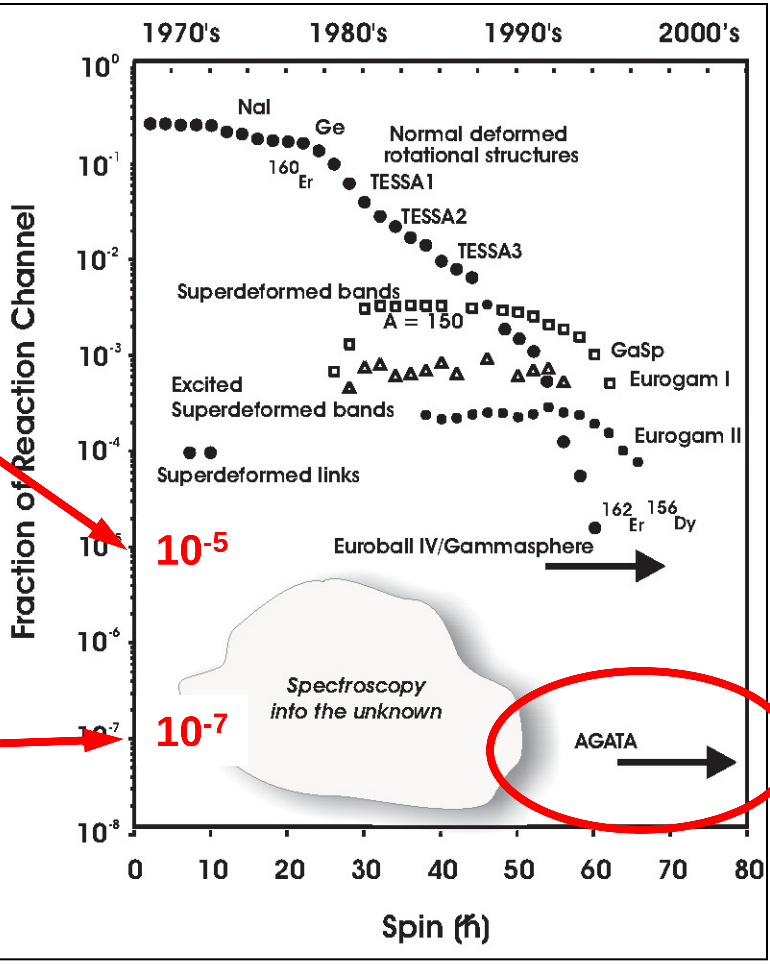
Ge Shielding taken away
Ge electrically 36-fold segmented

γ-Ray TRACKING technique
in electrically segmented Ge.

Involves:

- Digital sampling electronics
- Algorithms to extract time, position and energy, from **P**ulse **S**hape **A**nalysis (PSA)





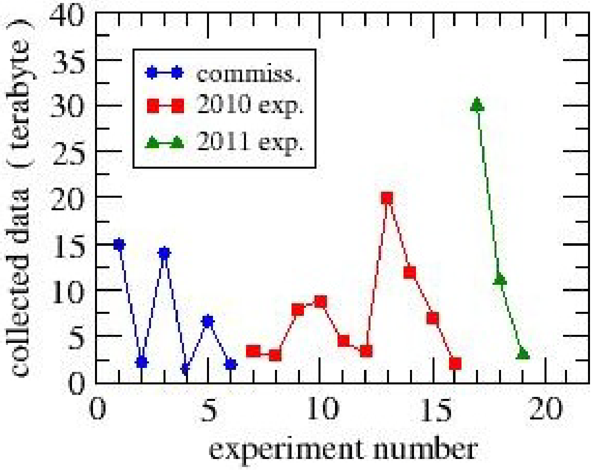
MIGRATION TOWARDS THE GRID

Previous Generation:
GAMMASPHERE, EUROBALL

- Less than 300 GB data per experiment
- Less than hundred sequential files to process
- Data stored on Exabyte tapes and analyzed at home institutes

New Generation: AGATA (PHASE-I)

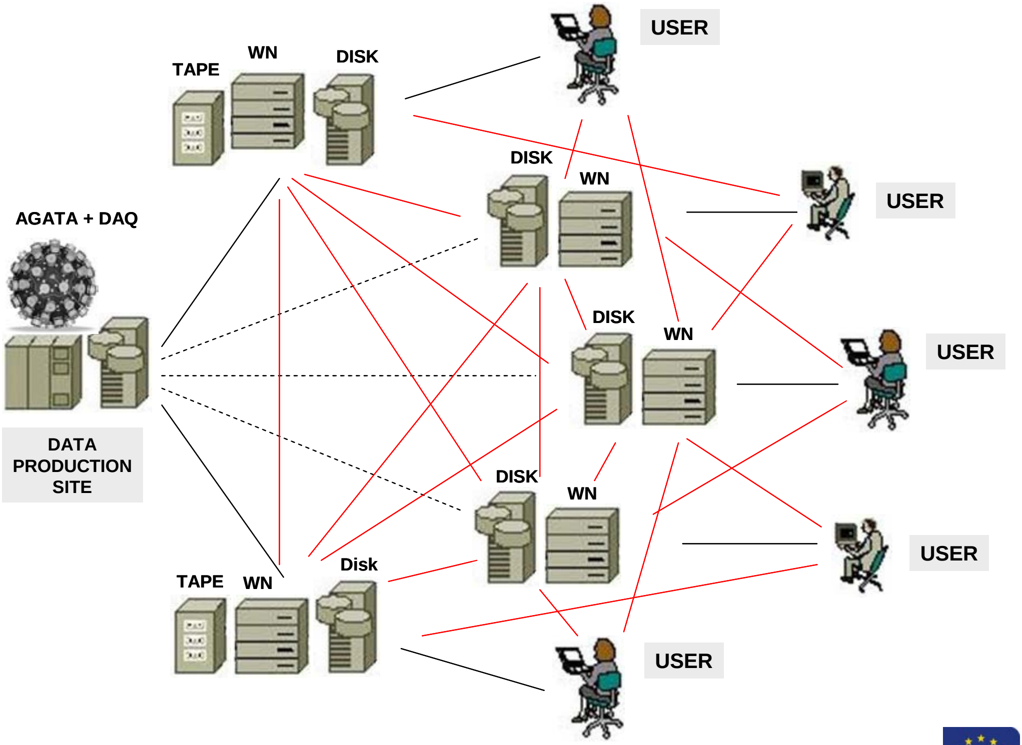
- More than 157 TB of Raw Data produced
- Around 8-10 TB per experiment
- Thousands of files 3-5 GB each to process
- Slow Data (re)Processing on a single CPU
- Migration to Grid Storage



A new situation where the users share computing resources and data in a coordinated way (policy) within a Virtual Organization…

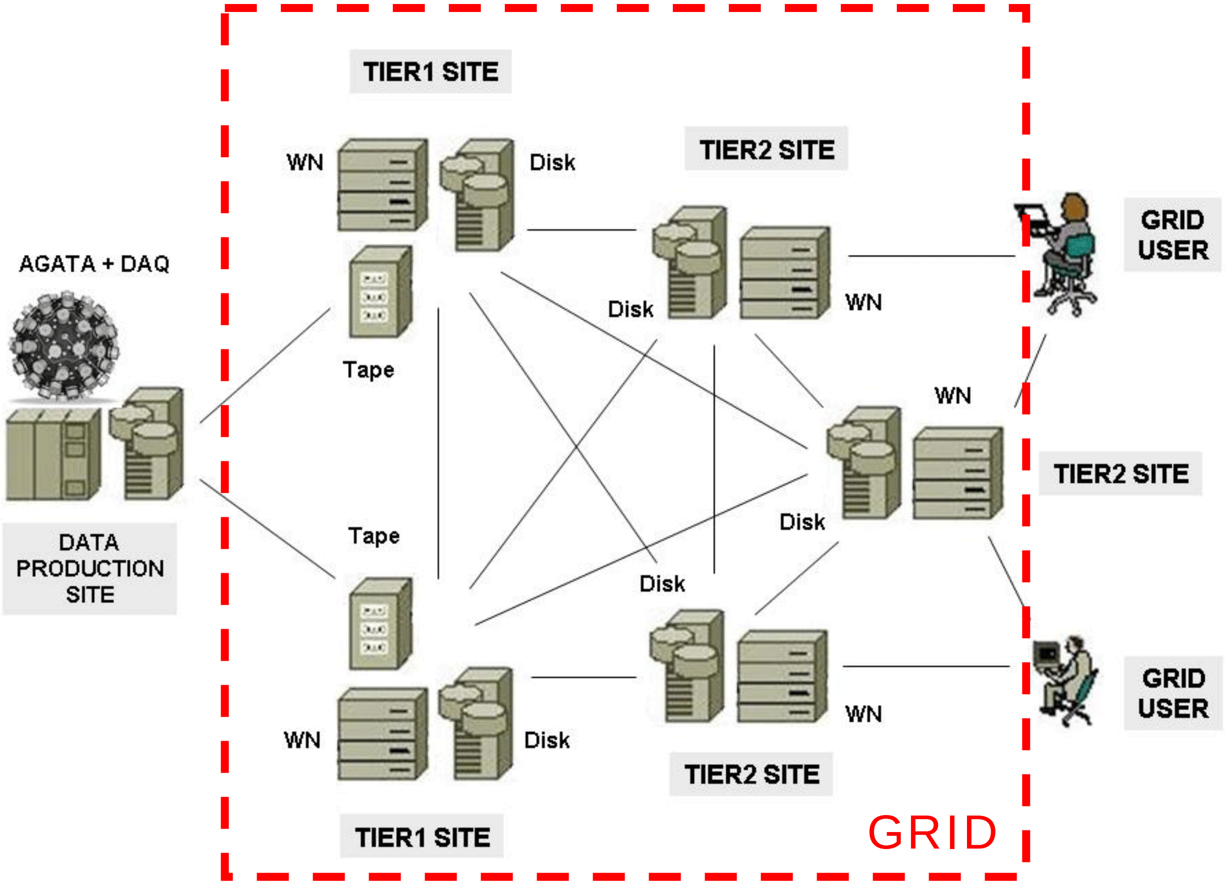
MIGRATE TOWARDS…

NEED A GRID COMPUTING MODEL…



THE AGATA GRID COMPUTING MODEL

The AGATA Grid Computing Model uses the LHC Tiers structure…
But, the Tier1s and Tier2s have similar roles in AGATA…
Tier1s provide Tape Storage…



DAQ: raw data, on disk (cached)

↓

Tier1: raw data, T1D0

↓

Tier1/Tier2: raw data, replica on disk (cached)

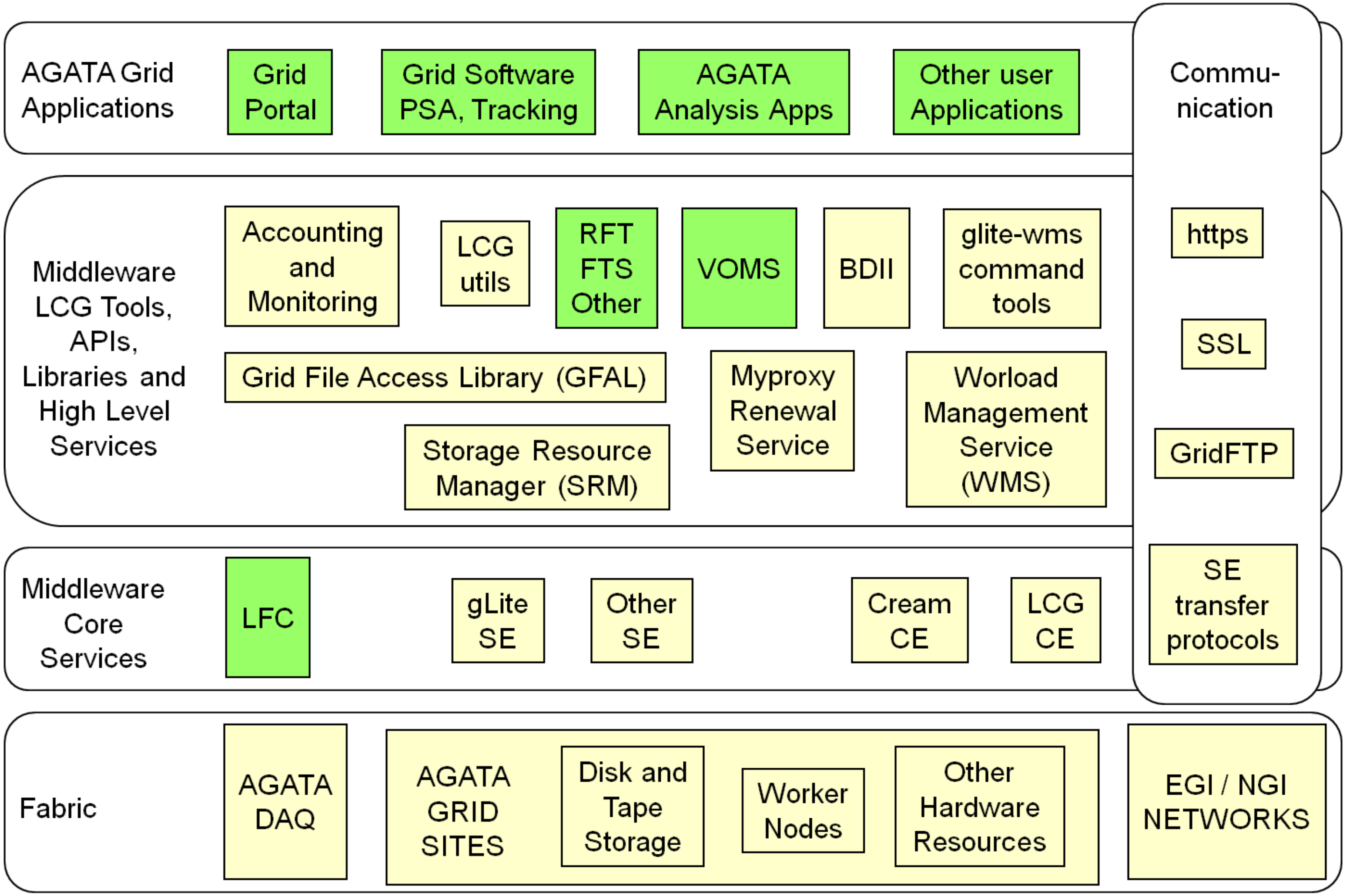
↓

Tier1/Tier2: PSA and Tracking raw data reprocessing. Production of reduced data for Analysis, T0D1

↓

User data analysis (ROOT)

THE AGATA GRID ARCHITECTURE



AGATA Grid Applications

Grid Portal

Grid Software PSA, Tracking

AGATA Analysis Apps

Other user Applications

Commu-nication

Middleware LCG Tools, APIs, Libraries and High Level Services

Accounting and Monitoring

LCG utils

RFT FTS Other

VOMS

BDII

glite-wms command tools

Grid File Access Library (GFAL)

Myproxy Renewal Service

Worload Management Service (WMS)

Storage Resource Manager (SRM)

https

SSL

GridFTP

Middleware Core Services

LFC

gLite SE

Other SE

Cream CE

LCG CE

SE transfer protocols

Fabric

AGATA DAQ

AGATA GRID SITES

Disk and Tape Storage

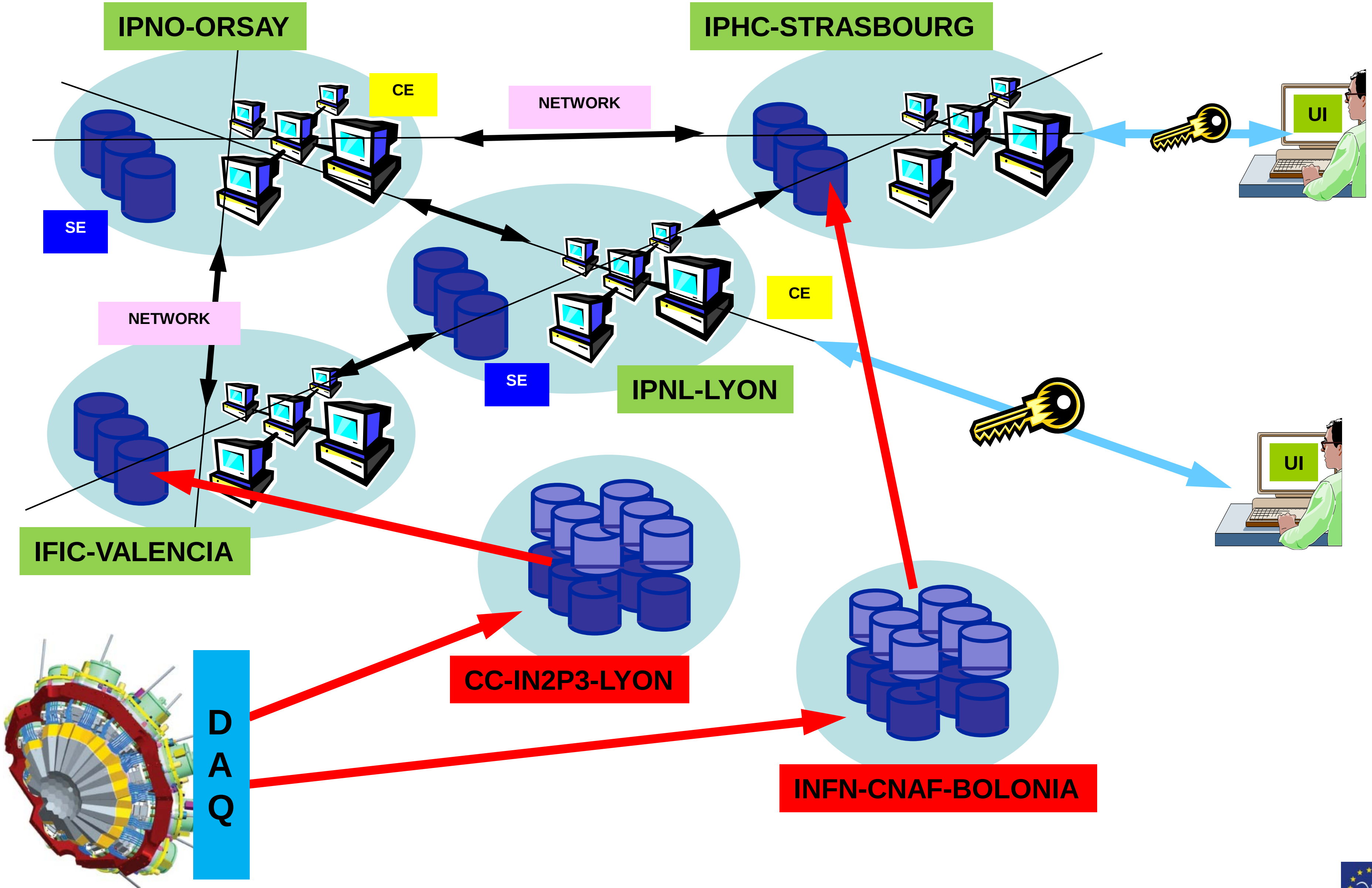
Worker Nodes

Other Hardware Resources

EGI / NGI NETWORKS

LEGEND: General EGI / NGI components Adapted components

AGATA GRID COMPUTING RESOURCES (I)



AGATA GRID COMPUTING RESOURCES (II)

◉ **AGATA**

Production site

🟨 **Tier-1 sites**

CC-IN2P3-LYON
CNAF-INFN-BOLOGNA

🟨 **Tier-2 sites**

IPNL – Lyon
IPHC – Strasbourg
IPNO – Orsay
IFIC – Valencia

● **Users**

----- **Data transfers**

— **Jobs**

AGATA GRID COMPUTING RESOURCES (III)

USER INTERFACE

lyoserv.in2p3.fr	*IPNL – Lyon*

COMPUTING ELEMENTS

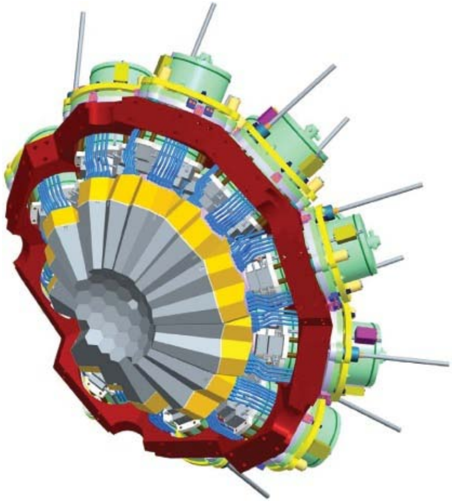
lyogrid07.in2p3.fr:8443/cream-pbs-vo.agata.org	*IPNL – Lyon*
ipngrid04.in2p3.fr:8443/cream-pbs-agata	*IPNO – Orsay*
sbgse2.in2p3.fr:8443/cream-pbs-vo.agata.org	*IPHC – Strasbourg*
ce03.ific.uv.es:8443/cream-pbs-agataL	*IFIC – Valencia*

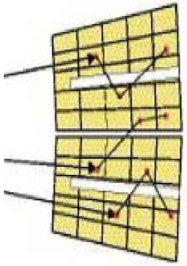
STORAGE ELEMENTS

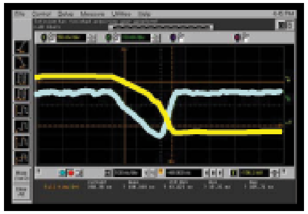
ccsrm02.in2p3.fr	*CC-IN2P3 – Lyon (Tier1 storage)*
storm-fe-archive.cr.cnaf.infn.it	*INFN-CNAF – Bolonia (Tier1 storage)*

lyogrid06.in2p3.fr	*IPNL – Lyon*
ipnsedpm.in2p3.fr	*IPNO – Orsay*
sbgse1.in2p3.fr	*IPHC – Strasbourg*
srmv2.ific.uv.es	*IFIC – Valencia*

AGATA DATA REPROCESSING ON THE GRID (I)







AVERAGE OF 10 TB DATA PER EXPERIMENT
AROUND 20 – 30 EXPERIMENTS A YEAR

DATA REPROCESSING BASED ON γ-RAY TRACKING TECHNIQUE IN ELECTRICALLY SEGMENTED Ge

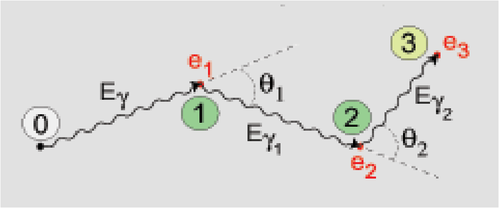
- Digital sampling electronics
- Algorithms to extract time, position and energy, from **P**ULSE **S**HAPE **A**NALYSIS (PSA)

Off-line PSA and γ-ray TRACKING reprocessing on the GRID: Software Migration

Pulse Shape Analysis to decompose recorded waves: **PSA Algorithm**

Identified Interaction **(x, y, z, E, t)**

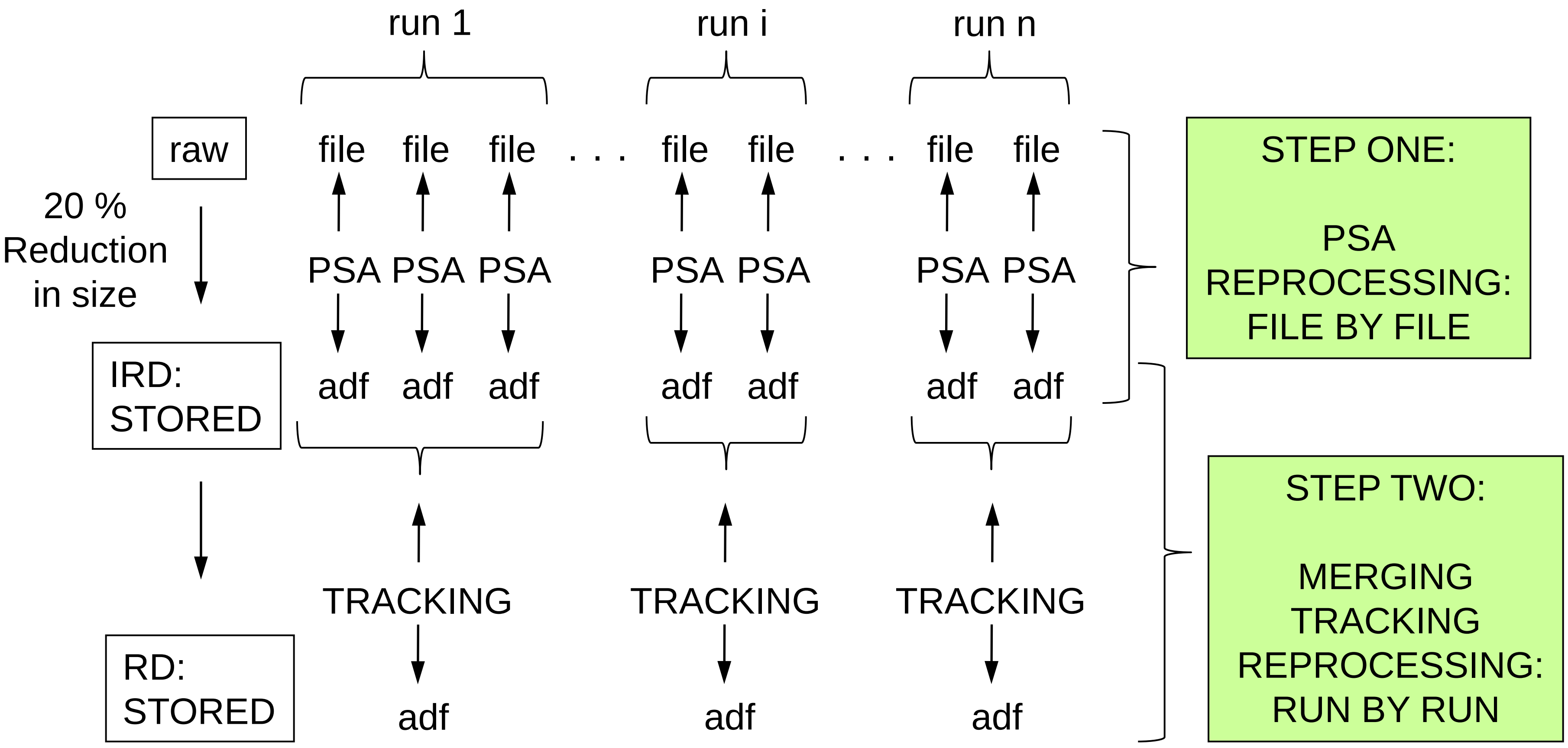
Reconstruction of tracks: **γ-ray Tracking Algorithm**



Recons-tructed γ-ray

AGATA DATA REPROCESSING ON THE GRID (II)

DATA REPROCESSING METHOD



run : Period of time where a given number of files is aquired and stored

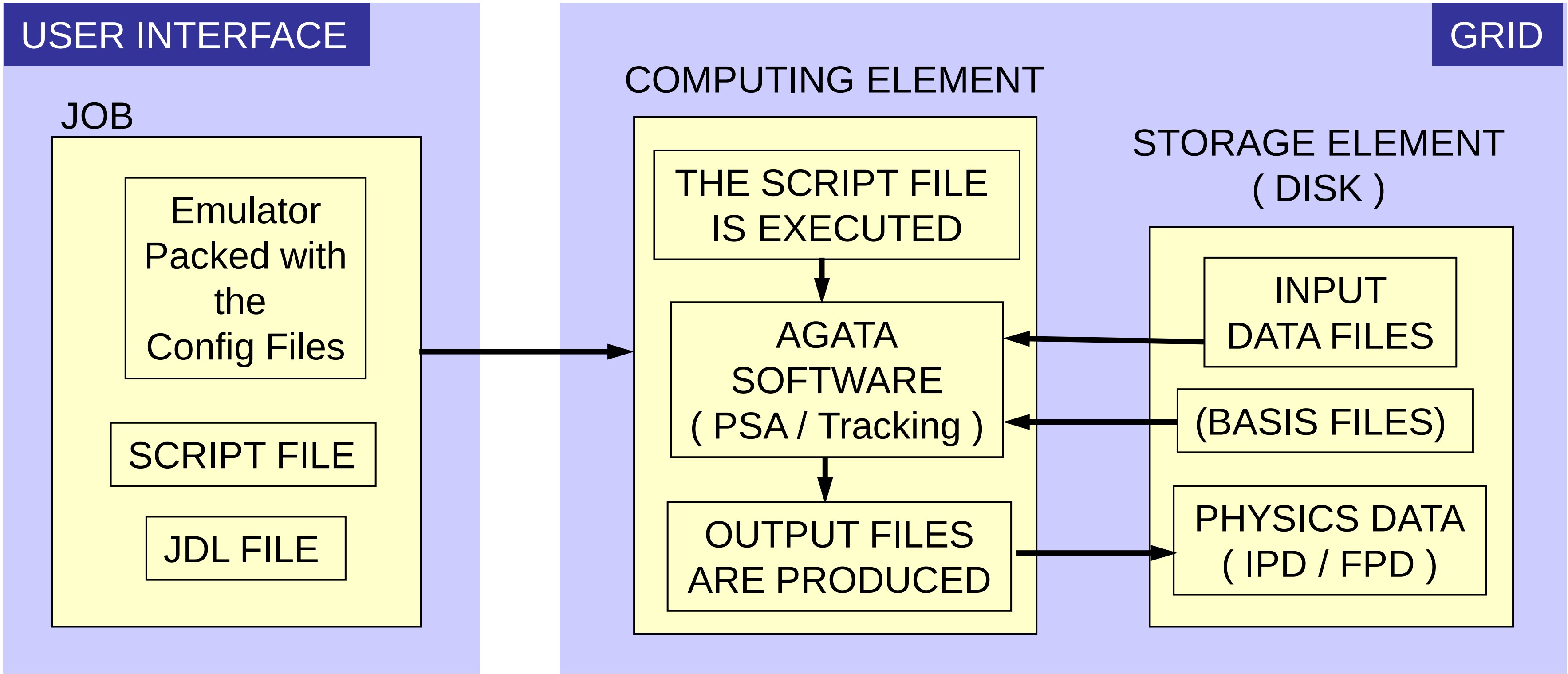
IRD : Intermediate Reduced Data

RD : Reduced Data, this is the Data to be analysed

adf : agata data format

AGATA DATA REPROCESSING ON THE GRID (III)

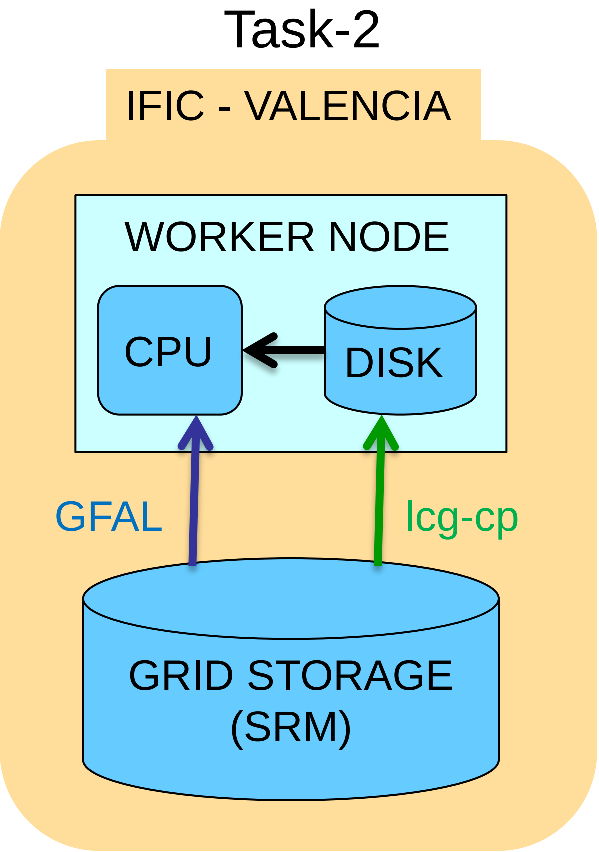
- prepare the configuration files for the Job
- define the configuration of the topology you want to use
- pack all the information with the emulator (here femul)
- create the script file to be executed in the WN
- create the JDL file describing your Job

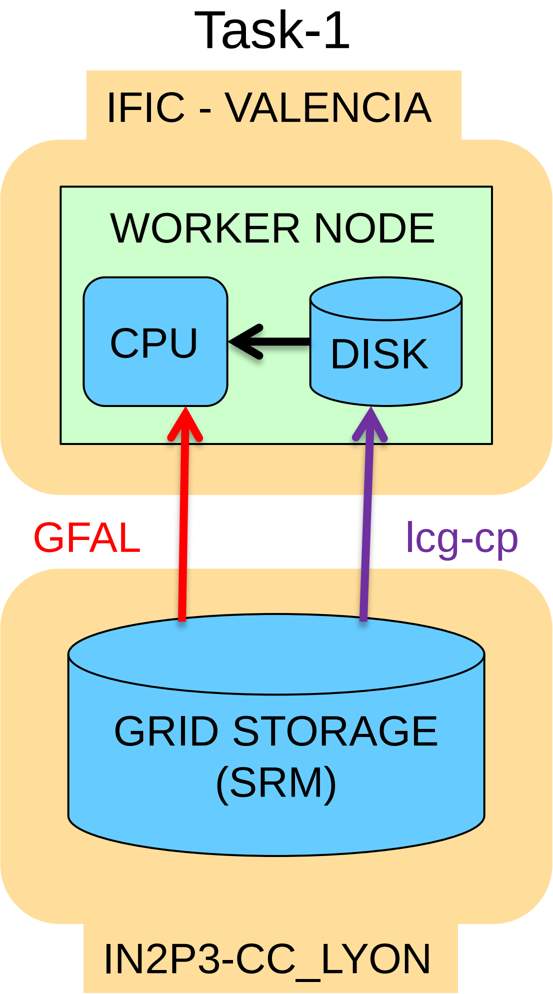


AGATA DATA REPROCESSING ON THE GRID (IV)

DATA ACCESS TESTS :
Local Access vs Remote Access using GFAL

- Two triple clusters (6Ge) 1B, 1G, 1R, 2B, 2G, 2R
- Each file is of 5 GB size
- Total of 30 GB processed
- PSA + γ-ray TRACKING processed
- A single Job submitted to the Grid





	Local Access	Remote Access (GFAL)
Task-1	56 ± 5 min	292 ± 32 min
Task-2	24 ± 4 min	28 ± 7 min

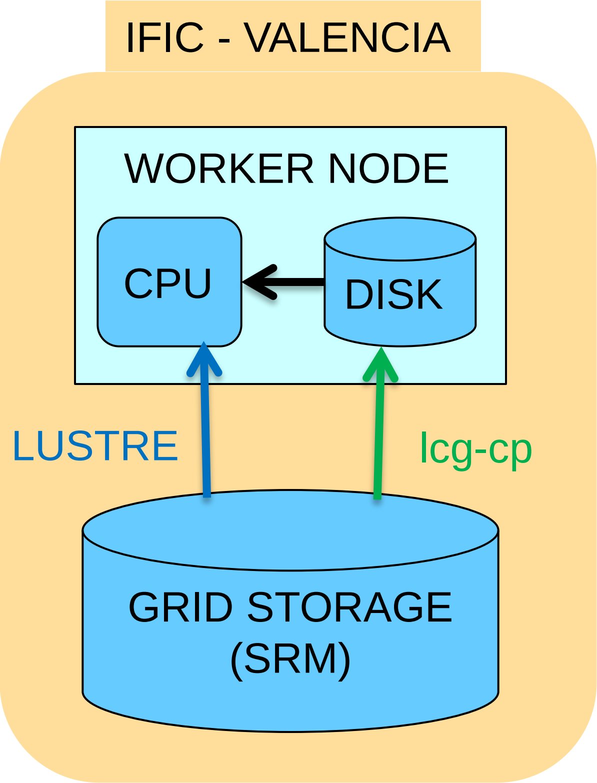
Running the Jobs where the data are located is highly recommended, using either the Local Access or the Remote Access method.

AGATA DATA REPROCESSING ON THE GRID (V)

DATA ACCESS TESTS :
Local Access vs Remote Access using LUSTRE

- 15 Ge crystals
- 30 files of 3 to 5 GB size each
- Total of 93 GB processed
- Only PSA is processed
- Task of 30 jobs defined and submitted to the Grid

	Local Access	Remote Access (Lustre)
only PSA	63 ± 4 min	52 ± 4 min



- 15 Ge crystals
- 373 Files corresponding to 25 Runs
- Task of 373 jobs submitted to the Grid for PSA
- Task of 25 jobs submitted to the Grid for γ-ray TRACKING

	Local Access	Remote Access (Lustre)
PSA	90 ± 9 min	73 ± 7 min
TRACKING	56 ± 5 min	38 ± 4 min

15% – 20% improvement in execution time when using Lustre compared to Local Access

AGATA DATA REPROCESSING ON THE GRID (VI)

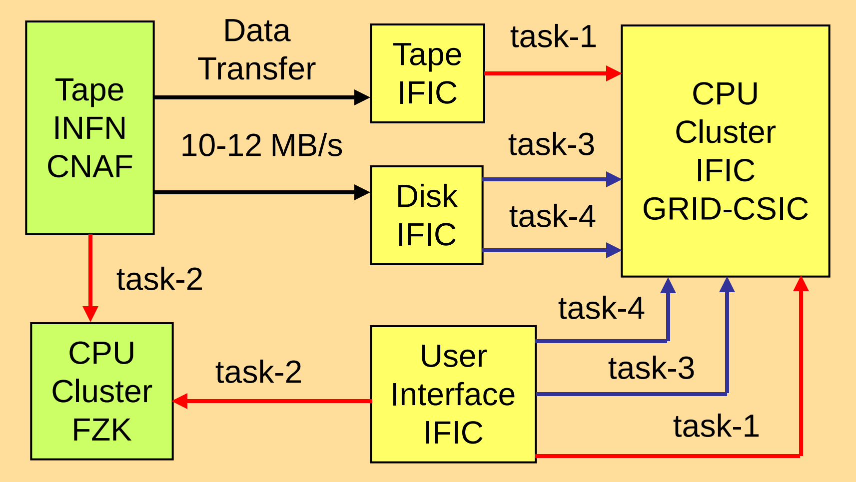
Old Work, from 2010

GRID RESOURCES USED :

GRID-CSIC : 50 cores (2GB per core, SLC5)
Sufficient Disk Space for storage (Lustre)

OTHERS : Additional storage 0.6 TB Tape (Castor)
Other EGEE clusters (CANF, FZK, MANCH.)

TESTS AND RESULTS :



Tape
INFN
CNAF

Data
Transfer

10-12 MB/s

Tape
IFIC

Disk
IFIC

task-1

task-3

task-4

CPU
Cluster
IFIC
GRID-CSIC

task-2

CPU
Cluster
FZK

task-2

User
Interface
IFIC

task-4

task-3

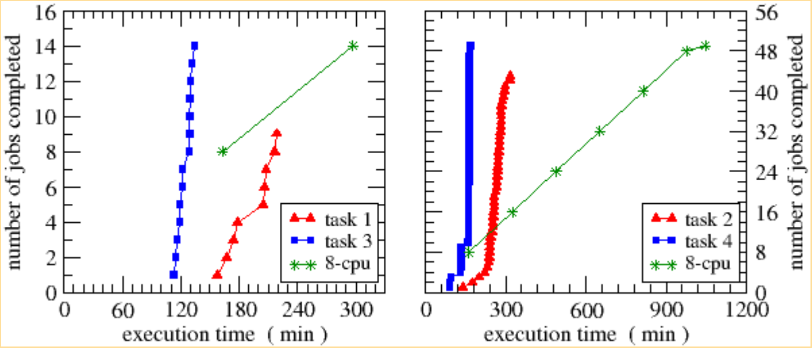
task-1

Task-1: 14 jobs, 0.6 TB data

Task-2: 40 jobs, 2.0 TB data

Task-3: 14 jobs, 0.6 TB data

Task-4: 49 jobs, 2.1 TB data



PRACTICE SESSIONS

HOW TO RUN A JOB ON THE GRID

I. Login into a User Interface machine and create a valid proxy

```
$> ssh –Y agastudX@lyoserv.in2p3.fr
$> voms-proxy-init –voms vo.agata.org
```

II. Write a JDL file and a Script file to be executed on the Grid

Content of a Script file:

- Uncompress the femul software with the right configuration
- Compile the software
- Download the Data files to be processed
- Run femul
- Upload the output and log files (if proceeds)

III. Grid commands for Job management

Submit the Job to the Grid :

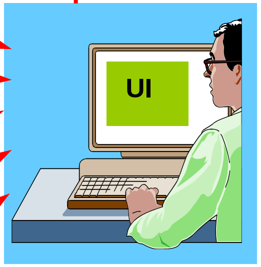
```
$> glite-wms-job-submit –a –o jobsID myJDLfile.jdl
```

Follow up the status of the Job:

```
$> glite-wms-job-status –i jobsID
```

Retrieve the outputs:

```
$> glite-wms-job-output –i jobsID
```



HOW TO RUN A JOB ON THE GRID

I. Login into a User Interface machine and create a valid proxy

```
$> ssh –Y agastudX@lyoserv.in2p3.fr
$> voms-proxy-init –voms vo.agata.org
```

II. Write a JDL file and a Script file to be executed on the Grid

Content of a Script file:

- Uncompress the femul software with the right configuration
- Compile the software
- Download the Data files to be processed
- Run femul
- Upload the output and log files (if proceeds)

III. Grid commands for Job management

Submit the Job to the Grid :

```
$> glite-wms-job-submit –a –o jobsID myJDLfile.jdl
```

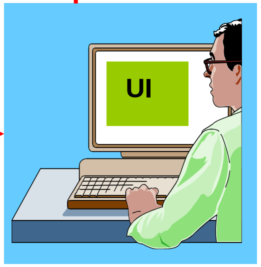
Follow up the status of the Job:

```
$> glite-wms-job-status –i jobsID
```

Retrieve the outputs:

```
$> glite-wms-job-output –i jobsID
```

USER APPLICATION



NAWAT-AGATA: HOW IT WORKS

THE USER PROVIDES:

In any case:

A Task Configuration File which contains the information about the task to be run on the Grid

```
InputStorage = srm://<storage-where-the-input-data-are-located>
OutputStorage = srm://<storage-where-to-upload-the-output-adf-files>
NumberOfJobs = <number-of-jobs-to-run-for-this-task>
ProcessType = <PSA-or-TR-or-ANC+TR>
DataAccess = <NONE-or-LUSTRE-or-GFAL>
```

A ConfExp/ directory that contains the configuration files for each run: Conf-run_xx/

A compressed copy of the femul software (to be installed on the Grid)

In case of PSA processing:

A file that contains the list of mappings Ge/BaseFile

A file that contains the list of the input data filenames (event_mezzdata) to be processed

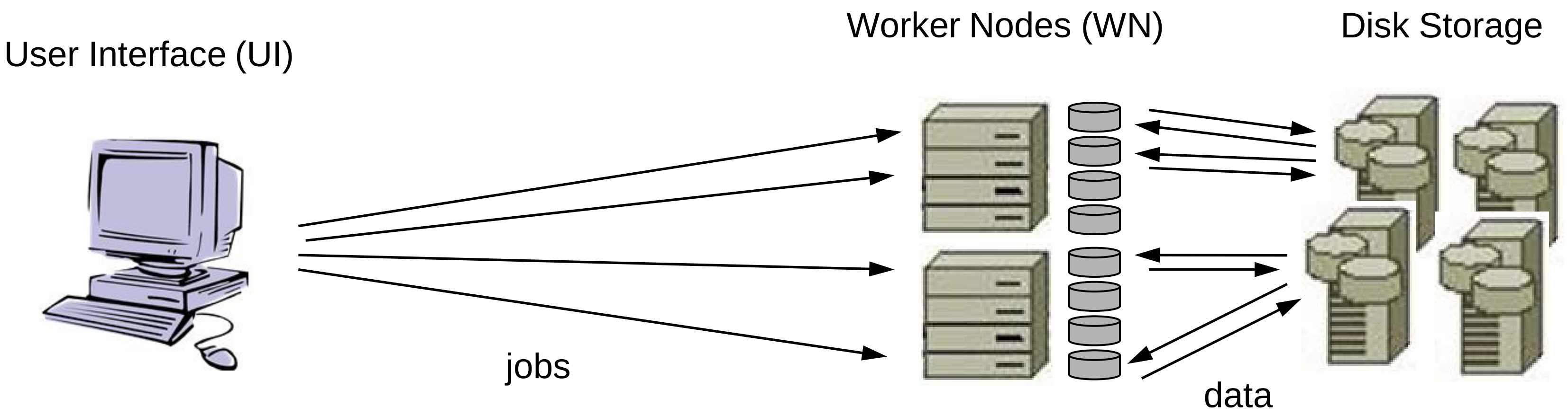
In case of MERGING/TRACKING processing:

A file that contains the list of the PSA_xx.adf filenames and Ancillary files (vmedata) if proceeds

THEN:

- Launch the nawat-agata application
- Select the Grid CPU resources to be used (Computing Element)
- Click the Execute button
- Go for a coffe or do some other work…

NAWAT-AGATA: HOW IT WORKS



Nawat-agata runs on the UI:

Generate the Script file
Generate the JDL file
Generate and compress the Config/
Submit jobs to the Grid
Follow up their execution
Retrieve outputs when jobs done

Jobs run on the Grid:

Uncompress software
Uncompress Config/
Download data files, if proceeds
Compile software (femul)
Generate Topology
Update BasicAFP/C
Run femul
Upload obtained adf files

Various Instances of femul running simultaneously on the Grid , each instance processing part of the data

ALL RUNS AUTOMATICALLY UNTILL THE TASK IS DONE