Road map for AGATA Computing model

O. Stézowski & E. Clément with the AGATA collaboration

The Advanced GAmma Tracking Array (AGATA) is a major European project, involving over 40 institutes in 12 countries, to develop and operate a high-resolution gamma-ray tracking spectrometer. Gamma-ray tracking requires the accurate determination of the energy, time and position of every interaction as a gamma ray deposits its energy within the detector volume. This is achieved by using electrically segmented hyper pure germanium detectors, pulse shape analysis of the digitised signals, and tracking algorithms to reconstruct the full event. The AGATA 4π geometry comprises 180 tapered-hexagonal coaxial detectors. AGATA can measure gamma rays from 10's of keV to 10 MeV with excellent efficiency and position resolution and has a very high count rate capability. These features result in an instrument with a resolving power of two orders of magnitude larger than previous spectrometers, such as EUROBALL in Europe and Gammasphere in the USA. The heart of the sensitivity of AGATA is the possibility to extract, from the Pulse Shape of the signals induced by the photon interacting in the HPGe crystal, the position of the impact. This position sensitive capability is currently obtained from the comparison on an event by event basis of the digitised signals of the 36 segments and the central contact and a reference basis. The set of 37 traces is computed on-line by a computer farm located at the host laboratory site where AGATA is installed. In Phase 1, the raw traces were recorded for a full reprocessing of the data to enhance the quality of the Pulse Shape Analysis and the Tracking algorithm to finally achieved the best quality of the γ-spectrum. With the linear increase of the available crystals in the AGATA array, recording the raw traces and later reprocess the full set of data will become a challenge which require to change the computing model in phase 2. These improvements should also pave the path towards open data models and open science approaches.

1 General Layout

1.1 data pipelines

Figure 1 gives the main ingredients of the data pipeline as realized in to-days $\gamma - ray$ trackers. The numbers given inside the pipeline are typical for a system running at $1kHz$ validation. Even if it could be possible to track inside a single Ge crystal, obviously, best performances (efficiency and P/T) are realized at the global level i.e. including all the Ge crystals in coincidence at a given time.

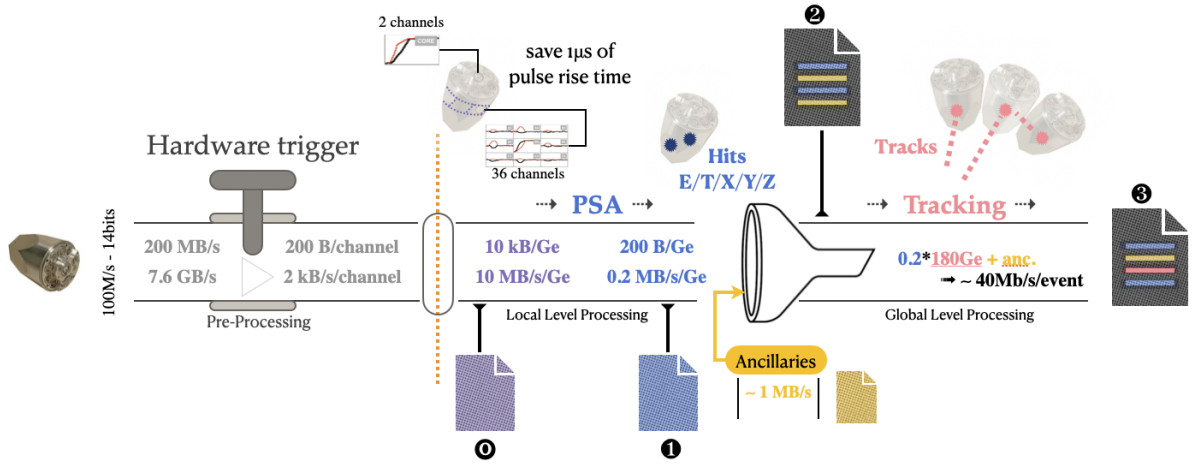

Figure 1: The data pipeline from left to right. 38 (2 cores + 36 segments) digitised signals from one Ge capsule are processed by the pre-processing electronic card (left side of the vertical, red, dashed line). Once something was detected in the crystal and the input is validated by the hardware trigger, samples over $1\mu s$ of the digitized signals are sent to a computing node (one per capsule for AGATA phase 0 and phase1) to achieve Pulse Shape Analysis and then to extract 3D interaction points. At the global level, coincidences, including ancillary data if available, are built before the reconstruction (tracking algorithm) of all the $\gamma - ray$ traces from the 3D hits. Produced data can be saved at various locations along the pipeline, from 0 to 3.

To implement such pipelines, the AGATA array is equipped at the host lab of the needed infrastructure to collect the in-beam data at high bandwidth. Presently, each crystal sends its data to a dedicated HTC node (through PCIe point to point connection) which collect the raw traces into the DCOD data acquisition, prepare the data for the Pulse Shape Analysis, run the Pulse shape analysis and, in phase 1, pushes both the raw traces and results of the PSA to the AGATA local storage (see in figure 1, collection points 0 (purple file) and 1 (blue file)).

The AGATA storage infrastructure utilizes the CEPH cluster technology. It employs a distributed storage architecture consisting of four disk nodes, housing a total of 64 disk units, providing a massive storage capacity of 466 TB. This storage system is commonly referred to as the AGATA Tier0. To ensure data integrity and prevent loss, the online disk server employs a triple redundancy mechanism, which allows for $\sim$ 130 TB of available space for data taking. The objective for AGATA in a 4π configuration is to reach 1Po of available storage.

The local infrastructure installed at the host laboratory consists also in workstations used for the local administration of the DAQ system, access, DCOD system and data flow etc ... Several workstations are also used for the on-line and near-line analysis of the collected data.

On can define specific points along the AGATA data processing pipeline (0 to 3 in Figure 1 and

Level	Format	Size (/s/det/kHz)	Content	(On/off)line	Software	Resp.
0	cdat	4 MB	Traces/det.	On	DCOD	Collab
1	adf	0.2 MB	Hits/det.	On	DCOD	Collab
				Off	*Replay*	*Collab-P.I*
2	adf	0.2 MB	Hits+anc.	Off	Replay	Collab-P.I
				On	*DCOD*	*Collab*
3	root	$\leq$0.5 MB	Tracks+hits+anc.	Off	Replay	Collab-P.I
				On	*DCOD*	*Collab*
4	root		Final (Conditions)	Off	Replay	P.I

Table 1: Online=in-beam,real time processing, offline = on site during the data taking or at home. DCOD = AGATA Online Data Acquisition, replay = femul software (local infrastructure, P.I infrastructure or GRID). Collab = AMB + Host lab team, P.I = experiment spokesperson team. If two lines for a given level, the first one corresponds to the most current common practice. In gray are the data exclusively produced out of the TIER0 site.

0 to 4 in table 1). At each point, table 1 provides some information on the data produced. The level 0 is the raw level containing the raw traces from the detectors and calculated energies, time and time-stamps as performed in real-time by the pre-processing card. The volume is 4 MB/s/ per AGATA capsules per kHz validated by the hardware trigger. The data are produced on-line using the AGATA DCOD data acquisition. The collaboration is responsible of the level 0. The level 1 corresponds to the Pulse Shape Analysis output containing only the hit position and energies for each AGATA crystal. Level 1 is performed on-line but can be redone offline from level 0 if recorded. Level 1 is always recorded on-line as Level 0 can be not saved in case of high rate.

At the level 2, the software performs the global merging of the data, i.e. all available AGATA crystal and complementary instruments based on the time-stamp. Level 2 is usually performed near-line during the experiment using the analysis servers and offline. Level 3 consists in the production of the final ROOT:Tree with full calibration and processing of AGATA (Tracking included) and analysis of the ancillaries. Level 4 is the final result after selection. The spokesperson team is responsible of level 4. Level 1, 2 and 3 are a close collaboration between the AGATA collaboration and the spokesperson team.

1.2 data management

After an experiment, the data are listed per spokesperson and copied to the CNAF (Bologna) and CCIN2P3 (Lyon) TIER1 as part of the AGATA VO on the GRID. The data recorded on the local CEPH disk are deleted when necessary after the copy to both TIER1 is secured.

After the irradiation beam time, the AGATA data are consistently transferred[1] in a ”duplicate mode,” ensuring that two separate Grid Tier1 storage computing centers receive copies for enhanced safety measures. This process is carried out following nearly every experiment to enable users to conveniently access and process the data. Concurrently, it effectively frees up the capacity of the current online CEPH storage, optimizing resource allocation for improved system performance.

[1]The current management of incoming and outgoing bandwidth involves the implementation of four distinct queues, each with a maximum limit of 300 MB/s. These queues operate in parallel and are dynamically managed to prioritize low latency, particularly during the execution of concurrent large data transfers.

Presently, the CCIN2P3 in Lyon and the CNAF in Bologna are the TIERS1 used by the collaboration.

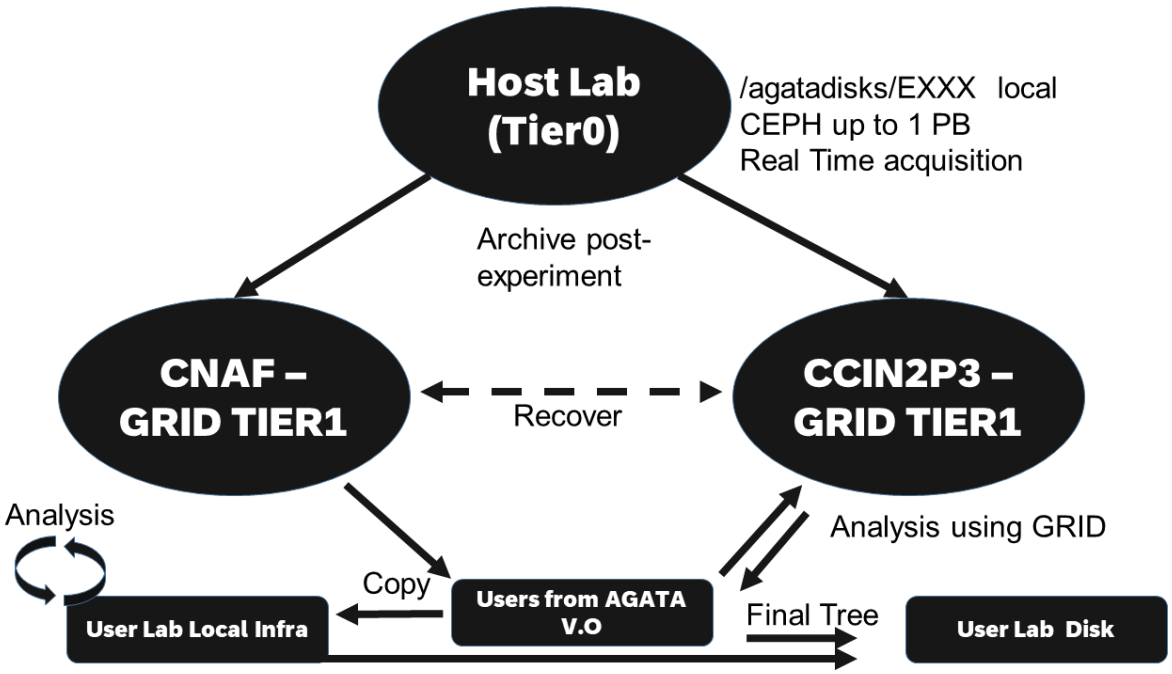

Figure 2: AGATA data work flow during Phase 1 [2010-2021] and during early Phase 2.

The spokesperson is advised of the copies and get the list of the data transferred. The AGATA collaboration archives the raw data produced (but do not keep track of the final processed data which belong to the PI team) together with some (minimum required to perform a replay) metadata on the GRID : it mainly consists on a raw copy of the online top directory containing many sub-directories and files.The AGATA collaboration admins the AGATA VO, provides the script to recover the data from the Tier1 and maintains a git repository of the data analysis codes, associated documentation, on-line forum for data analysis and organizes annually a data analysis school.

For the analysis, the PI team either processes directly on the GRID or, in most of the case copy, download from the GRID the data to its local infrastructure to analyse (see section 3). Figure 2 displays this work flow.

2 Tier 0 Tier 1 infrastructure for on-line; status and perspectives

At 1kHz validation, each core produces 8-10MB/sec with raw traces. After compression (cdat), raw traces and raw data represent 4-5 MB/sec at 1kHz validated. Without traces, i.e. keeping only the results of the PSA, the amount is reduced at 0.2 MB/sec/kHz. See for details figure 1. In Phase 1, until 2021, the AGATA collaboration has produced, keeping the traces, 900To which have been copied to the TIER 1.

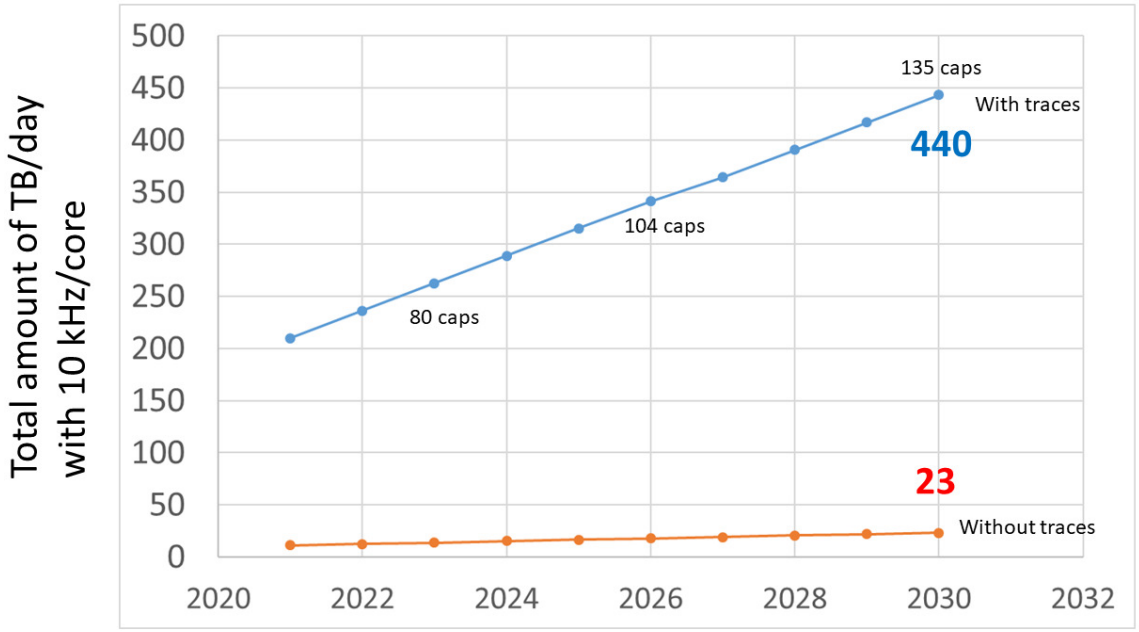

Figure 3: Daily amount of TB produced by AGATA at 10 kHz / core as function of years

The AGATA project definition defines a road map of capsules construction up to 135 crystal in 2030. One of the major improvement is the capability to run more than 10 kHz PSA per core. Using these inputs leads to the curves shown in figure 3. Keeping the traces would represent 440 TB per day or 23 TB per day without traces. At LNL, with roughly 70 days of data taking per year, it represents 1.6 PB per years without traces, and 30 PB with traces. This is obviously not doable.

The 10KHz PSA cannot be sustained with the present architecture. A closer look into the real condition is needed. First one should consider a real availability of crystal in the array, the type of experiment and finally the available beam time at the different facilities.

Based on the Phase 1 achievements, three types of experiments are defined

- Heavy Ions collision with magnetic spectrometer (PRISMA, VAMOS, MARA, RITU), validation rate $\sim$ 100 - 200 Hz / core
- Heavy Ions collision with particle tagger (NEDA, EUCLIDES, DIAMANT), validation rate $\sim$ 1000 - 4000 Hz / core
- Radioactive beams (MUGAST, GRIT), validation $\sim$ 100 - 200 Hz / core

The GANIL campaign between 2015 and 2021 covered the three cases and gives reliable

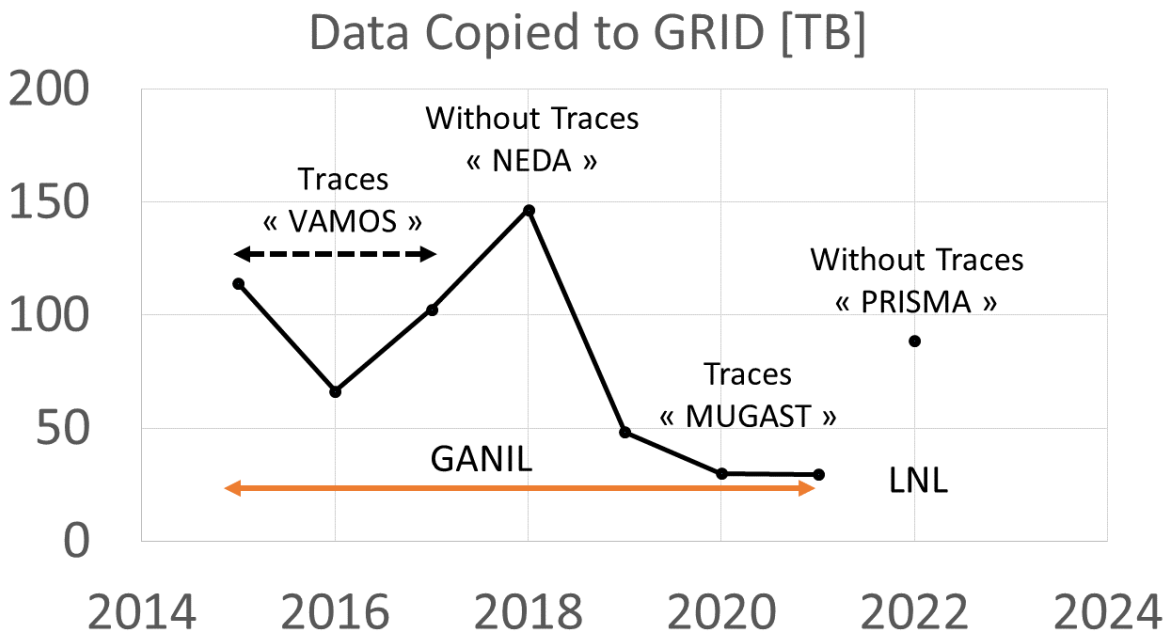

Figure 4: Annual TB copied to GRID TIER1 CCIN2P3 over the last decade

	2022	2023	2024	2025	2026	2027	2028	2029	2030
Crystal in Op.s	30	35	35	60	70	80	90	100	110
Exp.	PRISMA	PRISMA	PRISMA	NEDA	NEDA	SPES	SPES	FAIR	FAIR
Ind. Valid. [kHz]	0.2	0.2	0.2	4	4	0.2	0.2	0.2	0.2
TB per year	31	37	37	347	404	84	94	87	96

Table 2: 12 experiments per years and 72 days of data taking are considered in this estimation at LNL (PRISMA, NEDA and SPES). At GSI/FAIR, 20 days are considered. 2 TB per experiment per 30 crystals are recorded with Traces for calibration and qualification.

estimations. The annual volume copied to the TIER1 CCIN2P3 along the GANIL campaign is shown in figure 4 and allow to draw perspectives for the phase 2 adapting the estimation to the validation rate, days of beam on target and number of detectors in the array. Table 2 shows the estimated amount of data produced by AGATA until 2030.

Figure 5 summarizes the evolution as a function of the years without traces storage except for pre and post calibration runs. In average the expected volume is around $\sim$ 100 TB with a peak of production during the so-called NEDA campaign in 2025 and 2026 with an expected production of 300 to 400 TB per year. The present infrastructure of AGATA at LNL cannot, as it is, absorb such amount of data. The 2024 budget request will anticipate a significant increase of available storage at the TIER0 and encourage good and efficient practices to move the data and access the data for analysis to and from the TIER1 (see also the section 3). This includes also good practices in removing the data from the storage disk and avoiding unavailability of the system. One possible method would be to make use of the main and spear CEPH and alternate the access point to avoid colliding resources on the network and services.

In order to evaluate the effective cost of the archive on the CCIN2P3 and CNAF center, figure 6 shows the accumulated TB archived on TIER1. At the end of the MoU phase 2, it is anticipated a volume of $\sim$2PB at CNAF and CCIN2P3 each.

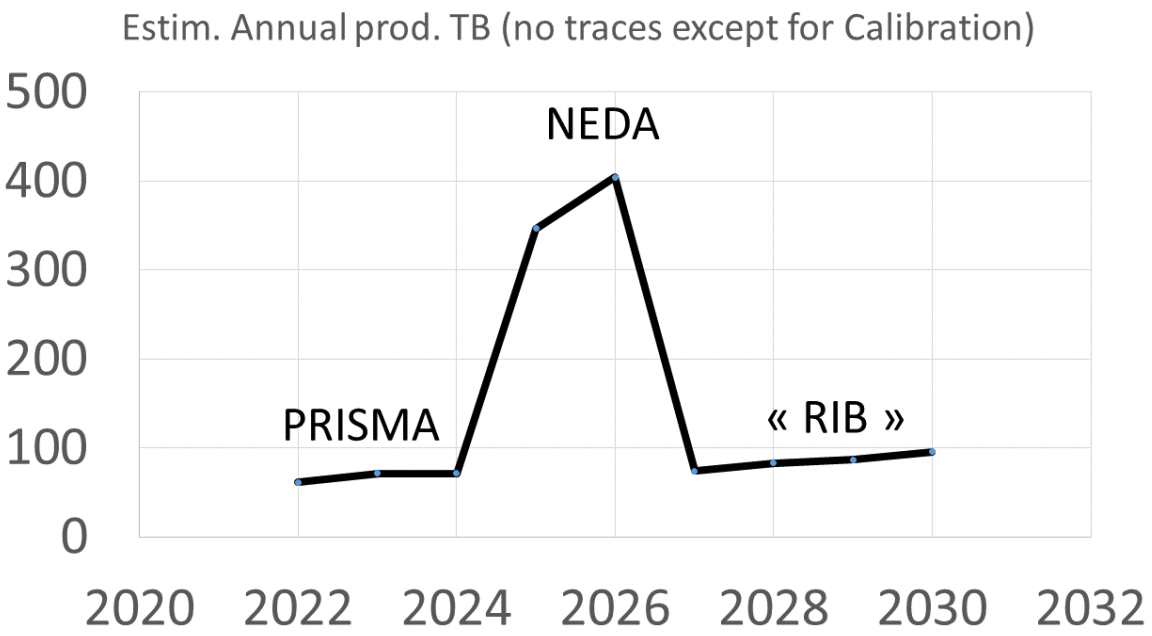

Figure 5: Annual TB to each GRID TIER1 over the next decade

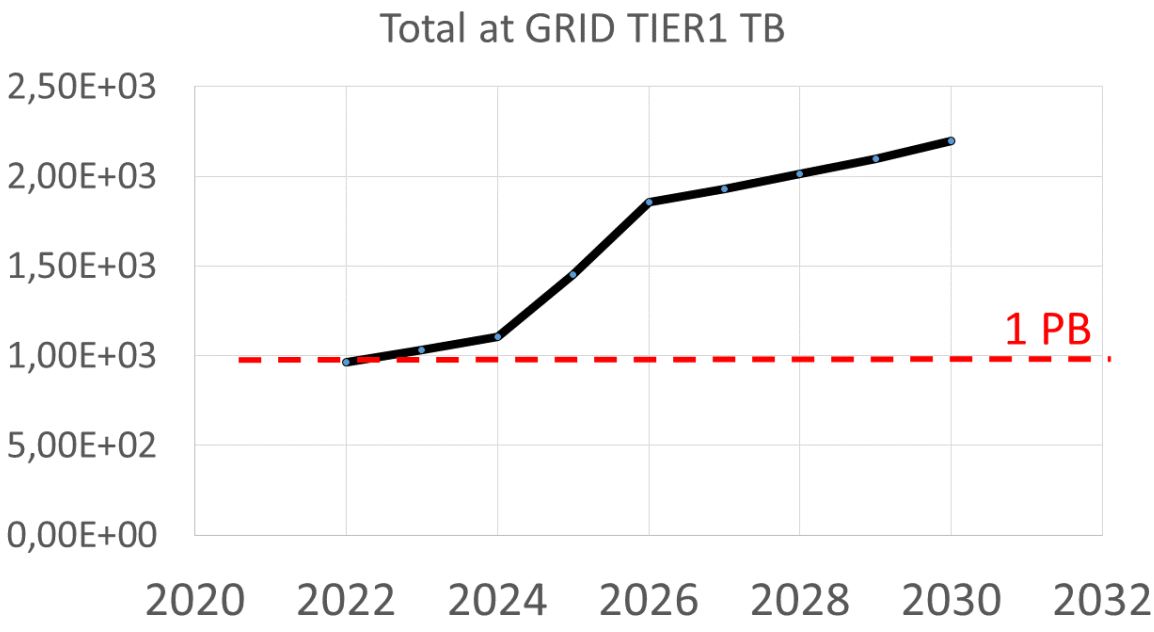

Figure 6: Accumulated TB to each GRID TIER1 over the next decade

3 Reprocessing : status and perspectives

The reprocessing of the data is part of the collaboration tasks and represent today a similar challenge as the data collection. They are of two aspects. The first one is the near-line analysis, the offline analysis and the long term sustainability towards Open access based on FAIR principles. The later is discussed in section 4.

The near-line analysis is the analysis of the data during the data taking. It consists on data quality checks and scientific objectives achievement. The online monitoring is performed by the AgaSpy software Dudouet, J. (2018). The near-line analysis is performed using a set of several HTC hosted in the DAQ-box and running directly on the TIER0 disk. The output is saved on a dedicated zone of the DAQ-box not interfering with the data collection. For more details see Korichi, A. (2023), Stézowski, O. et al. (2023). No specific problems are anticipated on this aspect.

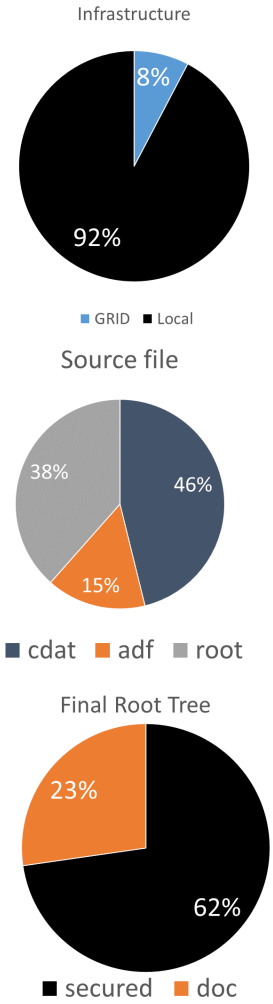

Figure 7: Survey of the usage in phase 1

The offline analysis is a separated problem since it is at the interface of the global effort of the collaboration during data taking and the transfer of the responsibility to the spokesperson team as described in table 1. During phase 1, the survey of the collaboration practices was done.

Figure 7 shows on which material the analysis was performed by the individual P.I team, from which level as defined by table 1 and how is stored the final analysis.

To date, the AGATA collaboration is responsible of the archive and is the owner of the raw data and analysis output performed near-line. The final analysis (level 4 in table 1) remains of the P.I team responsibility which store and describe the final analysed data set.

The survey shows that during phase 1 more than 90% of the data have been analysed with the local infrastructure of the P.I's team, nearly 50% from the PSA output (ie from level 1), 38% from the near-line RootTree (level 3) and only 15% from the raw data (level 0). In general, the output RootTree are secured at the P.I laboratory but rarely (23%) well documented.

The low use of the GRID is certainly linked to the balance between the needed processing time (1-2 weeks) and reasonable volume (max 50TB) per experiment, and the effort to access and bring the analysis at the GRID level. Teams which have analysed on the GRID often make use of a local expertise in the laboratory connected to HEP experiments. Others judges the balance more in favor of investing in a dedicated HTC with disk storage to run the analysis. It is worth mentioning that the level 0 to level 3 processing is done entirely (ie all runs) only once per experiment in offline.

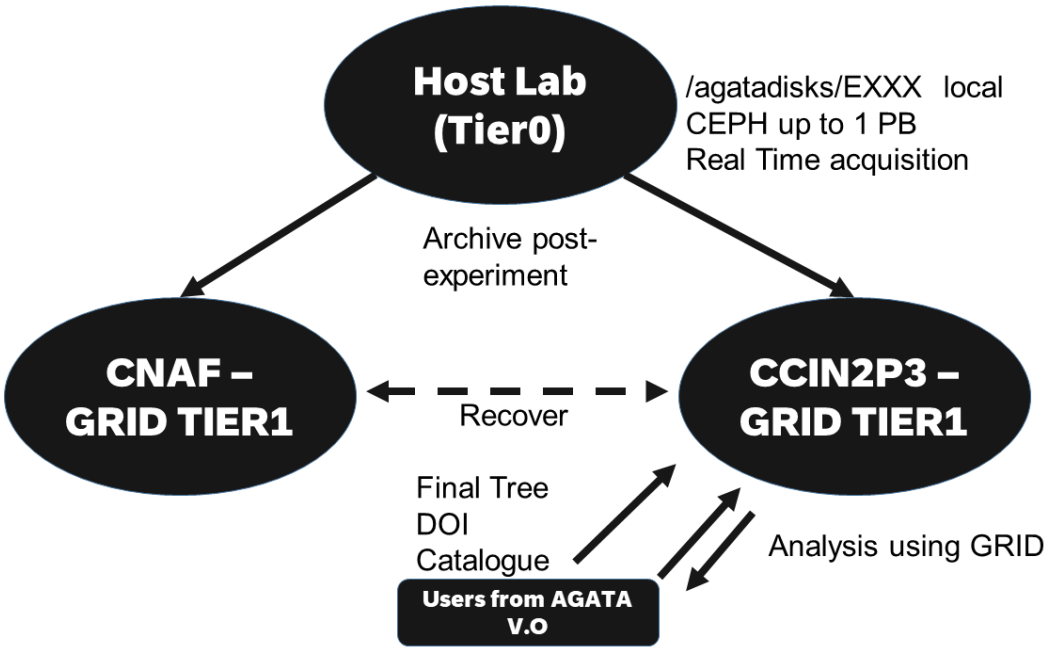

Figure 8: The new model ?

Since the second campaign in LNL (2021), the raw traces are only stored during the pre and post experiment calibration run's. A careful survey of the data quality is done during the beam time and traces are collected only to correct a major change of experimental condition. This goes in line with the practises observed during phase 1 and allow a limited impact on the collected amount of data even if the AGATA solid angle increases.

We however consider that some changes can be induced in the next phase. The procurement of many local HTC induces an increase of the needs for expertise and support from the collaboration expert to the local P.I team as a centralized system will simplified the maintenance and software deployment, and IT cost. The centralized system could be GRID or another system. The EOSC initiative promotes the use of data lake which could be an opportunity to make the remote access to the data easier. It promotes also the use of cloud computing but it is not clear for the moment if it matches the AGATA needs. The AGATA collaboration will keep use of a TIER1 infrastructure to archive its data. The centralized computing remains a question. In this

spirit, the software team is developing, based on DOCKER technology, a scalable approach for the code. Using the possibilities offered by the containers, the replay code could be used with a standard workstation, unique HTC or grape of HTC (cluster) at the P.I laboratory without intervention from the AGATA collaboration teams for support. Figure 8 displays this objective.

Finally its is the seed for a more FAIR and Open Access approach of the AGATA data. In particular the collaboration aims at improving its FAIR-ness approach as described in section 4.

4 Toward long term sustainability and Open access; status and perspectives

The current developments towards a new model are realized in order to :

- Make sure future PSA resulting from R&D can be integrated into the system. In particular, Machine Learning approaches may require to have GPU hardware used in data pipeline.
- Optimize the resources (RAM/disk storage/data transfer/processing time, etc ...) to handle up to 135 (180) Germanium crystals in the array (see previous sections) at higher rates
- Move to a satisfying open data and ultimately open science approach

The modifications should also follow the integration of the new electronic board realized in two steps. The first one is a cohabitation between the new electronics and the current one and then only the new electronics.

4.1 data pipelines

Figures 9 shows some of the foreseen modifications in terms of online processing workflow.

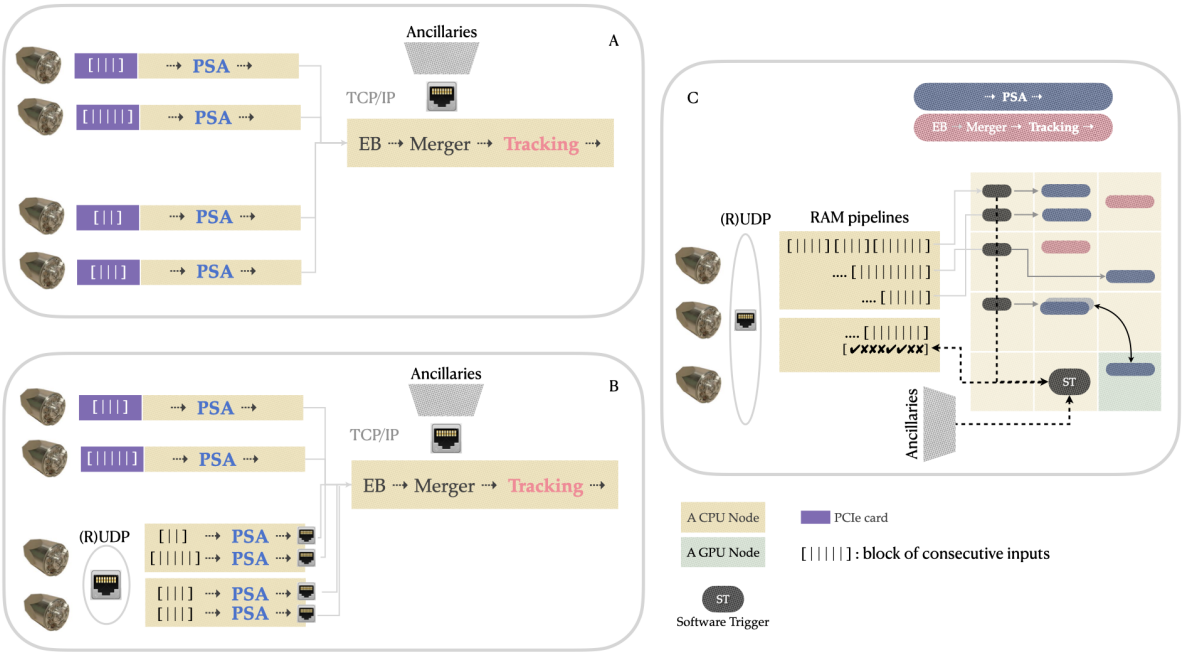

Figure 9: The new model ?

(A) This represents the current model. Traces are transferred from one crystal to a CPU node containing a dedicated PCIe card. The local level (i.e. PSA) pipeline is then performed in this node. The whole system thus relies on one computing node per crystal. The global level pipeline is executed also on a unique computer quasi-online/offline using the replay facilities or online through DCOD.

(B) This represents the foreseen approach for a cohabitation of the two electronics. For the new electronics, Traces are transported from the pre processing card to a computer node using the UDP (or RUDP) protocol. A daemon collects the UDP packets and pushes them in large buffers (one per crystal) in RAM. Connected to every FIFO, a DCOD chain performed the PSA as for the phase 0 and 1 electronics. What differs is the number of PSA that can run in a single node. Recent benchmarks performed on recent computing nodes show two current PSA algorithms can be run (possibly 3) at the same rate obtained for phase 0-1 local level pipelines.

(C) This represents the foreseen architecture for a system up to 4π. It consists first in having front machines in charge of buffering in large FIFOs the data coming from the pre-processing cards. To optimize the resources (CPU/network and disk), a Software Trigger is foreseen to, as fast as possible (before PSA is applied), identify at the global level only the most interesting events . In doing so, it should result in more interesting events in the data saved on disk. It should be noted that such software trigger (on multiplicity only) has already been tested successfully in DCOD environment. Another improvement is the use of dynamic load balancing either to parallelize a given pipeline (because of high counting rates in a sub-system) or to dispatch events in different algorithms. The best example for that is different PSA algorithms optimized for various situations corresponding to 1 or several hits in the same segment. Integration of heterogeneous hardware may also be crucial for future developments. Recent developments, using containerised processing nodes, have achieved the first step in that direction. With such architecture one can add as many computing nodes as required to operate at the desired (possible) rates. The amount of resources required (AGATA Phase 2 Project definition) has been so far estimated to 2 computing nodes per full-PSA i.e. PSA including treatment of 2 interactions per segments.

Since the beginning of AGATA, as much as possible, the various elements of the data pipeline are developed so that they can be used not only online in DCOD but also offline in Replay facilities. With the new phase, the same approach is to be continued and enhanced. With the usage of containerized processing applications, it is foreseen also to propose a new Replay facility able to break the limitation of the current proposed solution (monolithic application) which can effectively run only on a single computer. A first model has already been tested using docker (docker-compose) to run a replay on a single computer and docker-swarm on a cluster. It has also the potential to be used in more advanced, possibly cloud based, environment such as kubernetes. Whether or not this solution is to be deployed by users in their own local infrastructure or in a dedicated common center is to be discussed within the collaboration.

4.2 Data management

We identified major modifications in the data management at the production level but more importantly in the archive method and in the post-processing of the data. These modifications will be developed during the Phase 2. The AGATA collaboration has since the birth of the collaboration a Data Management Plan collaboration. It develops the ownership and embargo as well the publication rules. The Data Management commits the copy of the raw data with configuration files in the Tiers1.

Major modifications :

1. The very first modifications we would like to bring is a re organization and improvements of the meta-data linked to the raw data produced. It is our wish to save them the from tier0 to tiers1 and make them available to the P.I teams. The technological solution are based on the use of document database such as mongodb

2. In real time, we will re define the monitoring system and generalize the use of time serie

DB completing the meta data bases

3. Presently AGATA is keeping the raw data from the first experiment in 2010 (level 0 and level 1) and often the near-line level 2 and 3. There is no hierarchies in this data stored in the TIER1 which prevent us of for example, to delete useless or not state of the art or expired data. Therefore there is a need to re-organize and structure better the data archived in the Tiers1. We will develop the useof data catalogue and define for each level of the policy (licence/embargo/life time) and PID
4. A side product of the previous point is the need to recover from the P.I team the final level 3 and level 4 after carefull offline analysis. These processed data are presently out of control by the collaboration. These must be collected, included into the catalog and define for each level the policy (licence/embargo/life time) and PID
5. AGATA will rely on the long term on the TIERS1 infrastructures at CCIN2P3 and CNAF. Depending on the evolution of these facilities, there is the possibility of modification of the protocol (storage/access) to store/retrieve data to be taken into account by the collaboration.

AGATA is also progressing on the FAIRification

1. step 1 : new code developed keeping FAIR principles in memory i.e. configuration files and environment are json or xml based. It is should be easily read as 'machine' including a standardization of the words used as keys.
2. step 2 Upgrading the existing code to the standard.

5 Cost evaluation

The first cost in AGATA, in the particular item of computing and data management, is to secure a significant level of human resource in the teams involved in our data life-cycle. Data production, data storage, reprocessing and all activities related to the Open Data policy and advanced Data Management Plan require a dedicated team. Presently, the activity is supported by IP2i-Lyon, France, with two permanent physicists (60% FTE each) and one engineer (40% FTE) focused on management and software and by IJCLab-Orsay, France, with one permanent physicist (20% FTE) and four engineers (15% FTE each) for the IT infrastructure and data flow software. We consider this being the minimum.
Regarding the cost of the infrastructure, many alternative scenario can be drawn in term of topology depending of the physics request. We consider today that a scheme of 2 workstations per crystal is the proper envelop. The project definition detailed the associated investment budget for the period 2021 - 2030. We anticipate 916 k€ needed for the calculation nodes (HTC) procurement, 243 k€ for Tier0 data storage and 34 k€ for local analysis infrastructures.

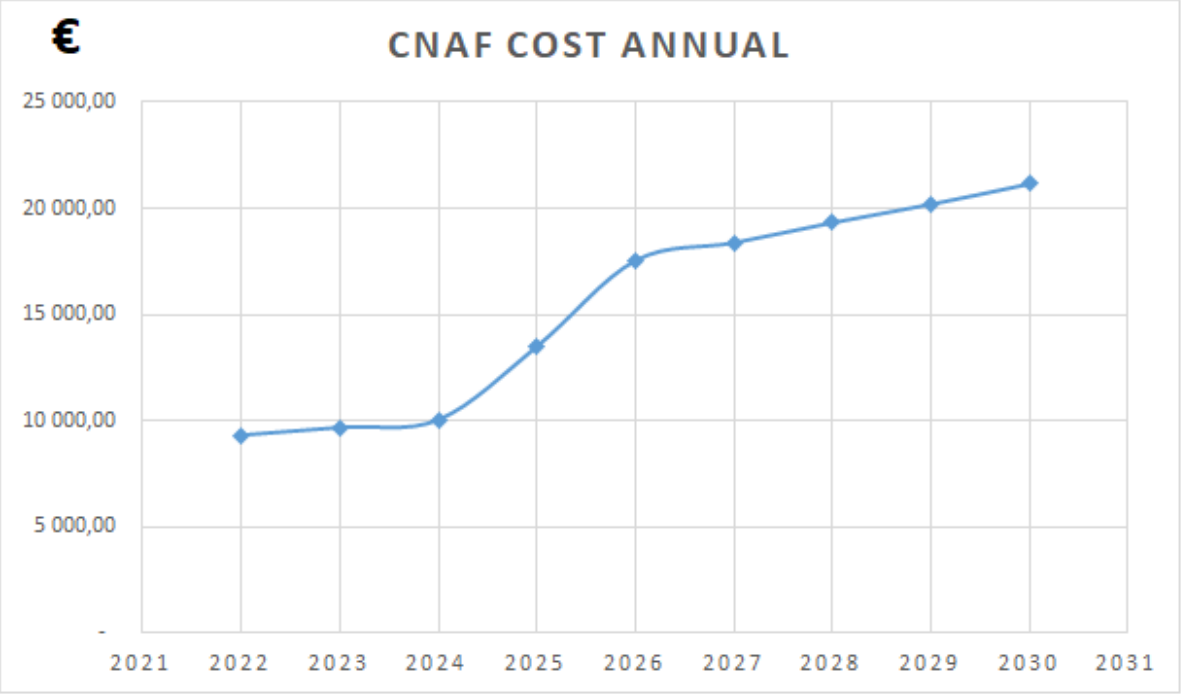

Figure 10: Estimated annual cost for CNAF Tier1

The archive to the TIER1 has a cost. Presently, CNAF charges the collaboration 10€ per TB stored on the infrastructure. The associated annual cost is shown in figure 10 as deduced from 6. At the end of MoU Phase 2, the annual associated cost for the storage in Tier1 is estimated to 20k€ to charged on the Operation Cost.

References

A. collaboration. Agata-data-policy. URL https://www.agata.org/acc/data_policy.

Dudouet, J. Agaspy: A new online data monitoring tool, 2018. URL https://indico.in2p3.fr/event/17160/contributions/64861/attachments/49914/63592/AG7-12_Dudouet_AGATAWeek2018_Agaspy.pdf.

Korichi, A. Agata daq-box: a unified data acquisition system for different experimental conditions. *Eur. Phys. J. A*, 59, 2023. doi: 10.1140/epja/. URL https://doi.org/10.1140/epja/.

Stézowski, O., Dudouet, J., Goasduff, A., Korichi, A., Aubert, Y., Balogh, M., Baulieu, G., Bazzacco, D., Brambilla, S., Brugnara, D., Dosme, N., Elloumi, S., Gauron, P., Grave, X., Jacob, J., Lafage, V., Lemasson, A., Legay, E., Le Jeannic, P., Ljungvall, J., Matta, A., Molina, R., Philippon, G., Sedlak, M., Taurigna-Quere, M., and Toniolo, N. Advancements in software developments. *Eur. Phys. J. A*, 59(5):119, 2023. doi: 10.1140/epja/s10050-023-01025-4. URL https://doi.org/10.1140/epja/s10050-023-01025-4.