PROCEEDINGS OF SCIENCE

Exploiting INFN-Cloud to implement a Cloud solution to support the CYGNO computing model

F.D. Amaro,[a] M. Antonacci,[b] E. Baracchini,[c,d] L. Benussi,[e] S. Bianco,[e] C. Capoccia,[e] M. Caponero,[e,f] D.S. Cardoso,[g] G. Cavoto,[h,i] D. Ciangottini,[j] A. Cortez,[c,d] I.A. Costa,[k,l,*] G. D'Imperio,[i] E. Dané,[e] G. Dho,[c,d] F. Di Giambattista,[c,d] E. Di Marco,[i] C. Duma,[m] F. Iacoangeli,[i] H.P. Lima Júnior,[g] G.S.P. Lopes,[n] G. Maccarrone,[e] R.D.P. Mano,[a] M. Marafini,[o] R.R. Marcelo Gregorio,[p] D.J.G. Marques,[c,d] G. Mazzitelli,[e] A.G. McLean,[p] A. Messina,[h,i] C.M.B. Monteiro,[a] R.A. Nobrega,[n] I.F. Pains,[n] E. Paoletti,[e] L. Passamonti,[e] S. Pelosi,[i] F. Petrucci,[k,l] S. Piacentini,[h,i] D. Piccolo,[e] D. Pierluigi,[e] D. Pinci,[i] A. Prajapati,[c,d] F. Renga,[i] A. Rodano,[e] R.J.C. Roque,[a] F. Rosatelli,[e] A. Russo,[e] G. Saviano,[e,q] D. Spiga,[j] N.J.C. Spooner,[p] S. Stanlio,[d] R. Tesauro,[e] S. Tomassini,[e] S. Torelli,[c,d] M. Tracolli[j] and J.M.F. dos Santos[a]

[a] *LIBPhys, Department of Physics, University of Coimbra, 3004-516 Coimbra, Portugal*
[b] *Istituto Nazionale di Fisica Nucleare, Sezione di Bari, 70126, Bari, Italy*
[c] *Gran Sasso Science Institute, 67100, L'Aquila, Italy*
[d] *Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali del Gran Sasso, 67100, Assergi, Italy*
[e] *Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali di Frascati, 00044, Frascati, Italy*
[f] *ENEA Centro Ricerche Frascati, 00044, Frascati, Italy*
[g] *Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro 22290-180, RJ, Brazil*
[h] *Dipartimento di Fisica, Università La Sapienza di Roma, 00185, Roma, Italy*
[i] *Istituto Nazionale di Fisica Nucleare, Sezione di Roma, 00185, Rome, Italy*
[j] *Istituto Nazionale di Fisica Nucleare, Sezione di Perugia, 06123, Perugia, Italy*
[k] *Dipartimento di Matematica e Fisica, Università Roma TRE, 00146, Roma, Italy*
[l] *Istituto Nazionale di Fisica Nucleare, Sezione di Roma Tre, 00146, Rome, Italy*
[m] *Istituto Nazionale di Fisica Nucleare, CNAF, 40127 , Bologna, Italy*
[n] *Universidade Federal de Juiz de Fora, Faculdade de Engenharia, 36036-900, Juiz de Fora, MG, Brasil*
[o] *Museo Storico della Fisica e Centro Studi e Ricerche "Enrico Fermi", Piazza del Viminale 1, 00184, Roma, Italy*
[p] *Department of Physics and Astronomy, University of Sheffield, Sheffield, S3 7RH, UK*
[q] *Dipartimento di Ingegneria Chimica, Materiali e Ambiente, Sapienza Università di Roma, 00185, Roma, Italy*

E-mail: igor.abritta@roma3.infn.it

*Speaker

PoS(ISGC2022)021

The aim of the CYGNO project is to demonstrate the capability of a high resolution gaseous TPC based on sCMOS (scientific CMOS) optical readout for present and future directional Dark Matter searches at low WIMP masses (1-10 GeV) down to and beyond the Neutrino Floor. CYGNO is a medium-size astroparticle physics experiment that requires a relatively small amount of computing resources and for this reason can be subjected to a fragmentation and low utilisation rate. This is a typical use case that could exploit and benefit from all the features of a Cloud infrastructure. In the context of the INFN Cloud project, a container-based system has been developed in order to provide a seamless integration between storage and computing system. The latter is based on JupyterHub to provide a multi user server to access the experiment environment (ROOT, GEANT, GARFIELD++, libraries, etc). The token based authentication and authorization system allows a seamless integration with S3 Cloud Storage where a remote DAQ system continuously uploads acquired files. The result is a "Software as a Service" (SaaS) layer for data analysis and simulation with common tools of our community. The paper will detail the overall project and preliminary user experiences.

International Symposium on Grids & Clouds 2022 (ISGC 2022)
21 - 25 March, 2022
*Online, Academia Sinica Computing Centre (ASGC), Taipei, Taiwan****

Contents

1. Introduction

The physics theories describing our Universe (General Relativity, the Standard Model of Particle Physics and the Cosmological Model [1]) are still full of questions and unknown issues. This scenario is manly investigated by the so called astroparticle experiments: hundreds of experiments located in many different environments such as deep underground laboratories, sea water, over highlands, on balloon and in the space. These experiments have to probe a plethora of phenomena that go from the Gravity and Gravitational dynamics effects, neutrinos physics, dark matter and dark energies to cosmic rays studies. Such experiments are of very different scale and the computer models they need to acquire, store, simulate and analyze data are very different and typically very far from High Energy Physics (HEP) experiments at accelerators (see Figure 1). They are producing "big data" with different volume, velocity and variety, and the reconstruction, analysis and simulation of most of them can not continuously be processed, producing a large discontinuity and sometimes inefficiency of the exploitation of the allocated computing resources. These motivations pushed us to study and implement the CYGNO use case on the INFN Cloud computing infrastructure.

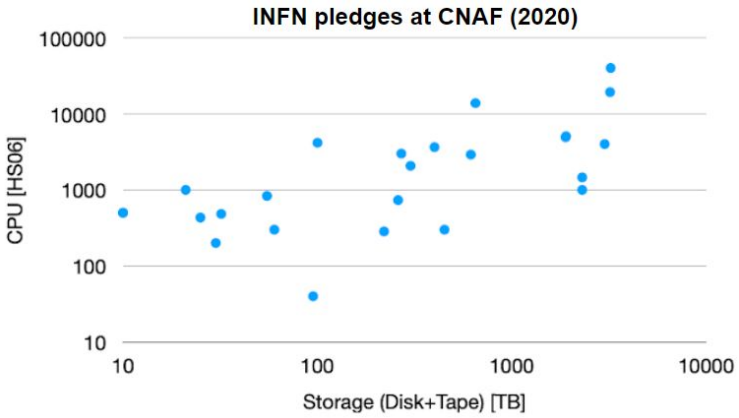

Figure 1: An example of the variability of requirements for computational resources (CPU vs Storage) requested by INFN astroparticle experiments at computer center CNAF (2020).

2. CYGNO Experiment

The aim of the CYGNO project [2] is to realize a large gaseous Time Projection Chamber (TPC) based on Optical Read Out (OPT) of the ionization electrons avalanche multiplications [3], exploiting the progress in commercial scientific Active Pixel Sensors (APS) based on Scientific CMOS (sCMOS).Those detectors can investigate O(1-10) GeV dark matter mass range with reasonable sensitivity, when reaching large volumes. It can reject beta/gamma background and exploit the feature of directionality, identifying DM/SN coming form a specific direction [4]. The roadmap (see Figure 2) foresees the construction of a N/m^3 demonstrator equipped with 18 cameras - APS sensors based, 2304×2304 resolution with single photon sensitivity - looking for the rare candidates events to be identified over beta/gamma background produced by natural radioactivity. The detector will be also equipped with 4 Photo Multiplier Tubes (PMT) symmetrically placed around each camera to detect the time shape longitudinal evolution of the detected tracks.

A first prototype, called LIME, equipped with a single camera and 4 PMTs has been installed at the beginning of the 2022 in "quiet" underground environment of the INFN Gran Sasso Laboratory (LNGS) and is starting to produce first data. The purpose of the project is to demonstrate that the technology is ready to build a O(30-100)m^3 detector.

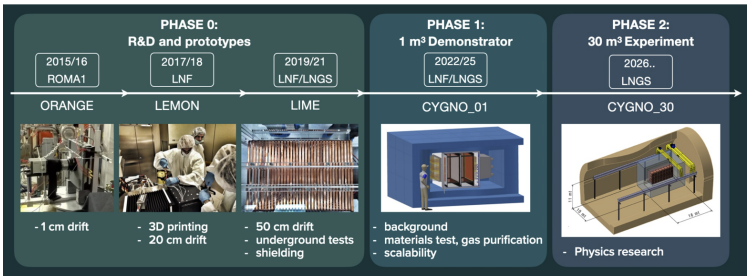

Figure 2: CYGNO experiment roadmap.

3. Computing requirements

The advantage of OPT based on APS sensors lies in the high sensitivity and granularity that such commercial sensors can offer with respect to standard electron amplifiers that could be very expensive, radioactive and with a limited density. This technology allows to build detectors with O(10^7) readout channels, achieving 100-150 μm space resolutions and $\simeq$10-12% energy resolution at eV scale, as it is foreseen for the CYGNO demonstrator (see Figure 3) with a cost that is O(10^3) lower than with standard electronics.

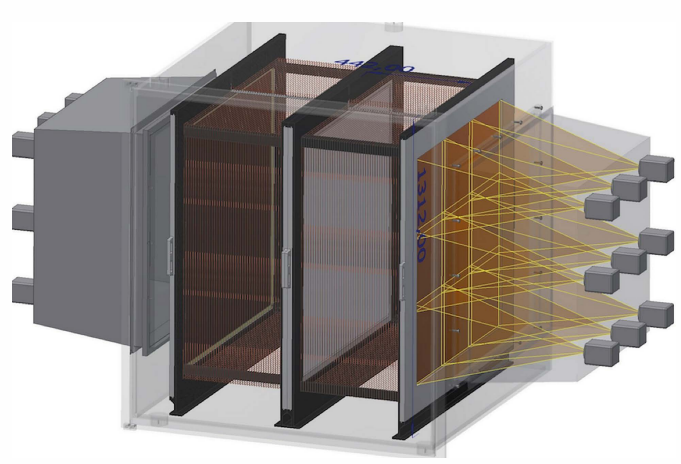

Figure 3: The 1m^3 CYGNO-phase1 demonstrator (CYGNO_1).

Obviously such technology is limited in speed and can be applied only in limited cases such as DM search and SN study where events rate and detector occupancy are very low. The prototype installed at LNGS today is producing 2.5MB of data size/event. Being the expected rate, due mainly to background to be rejected by the reconstruction process, about 0.2 Hz, this results in about 50 GB/day. Anyhow, the computing infrastructure has to be ready to acquire larger data throughput of CYGNO (about 20 times) and to demonstrate the feasibility of CYGNO30-100 (100-1000 times the current prototype). To the experiment data, simulation and calibration data have to be summed, and a total amount of 20TB storage/year is estimated for the prototype phase.

Even if CYGNO is a small/medium size astroparticle experiment from the data throughput perspective, reconstruction and simulation need to handle large images and clustering algorithms [5–7] are needed to identify candidate signals and reject background. Moreover, Machine Learning is largely used to identify sensor noise, particles direction and Monte Carlo validation. Last but not least the CYGNO project data and software need to be accessible and easy to use for all the international collaboration and students participating.

4. INFN Cloud project

Over the past two decades INFN has played a leading role in the design and implementation of large-scale computing infrastructures and applications for scientific communities. While initially this was primarily focused to meet the needs of the latest generations of HEP experiments, nowadays

it is rapidly extending to other communities. The national distributed system deployed in the context of the LHC Computing consists of 9 medium-sized centers (Tier-2) and 1 large Tier-1 center, at CNAF (Bologna). Moreover, at the beginning of 2020 the INFN Cloud project [8] has been launched with the aim of building a distributed Cloud infrastructure and providing advanced services for the INFN scientific communities. The INFN Cloud Infrastructure entered the production phase towards the end of 2020 and the project is now driving all the INFN initiatives related to the Cloud developments. The key features of the INFN Cloud can be summarized as follow:

- A multi-site federated Cloud infrastructure owned by INFN and possibly extendable to other Cloud infrastructures and resources.
- A set of services that can be used through a portal, from a terminal or with a set of APIs.
- A "high-level" mechanism for adapting and evolving the service portfolio according to the needs and requests of users.
- A fully distributed intra-INFN organization for the support and management of infrastructure and services.

On top of this the two key architectural elements are: a distributed resource orchestration and a modern federated solution for the identity access management. The new born INFN Cloud project resulted a suitable technology enabler for the CYGNO computing model implementation and to this end a series of R&D activities started.

5. Implementing CYGNO computing model on INFN Cloud

In order to exploit the INFN Cloud resources and service solutions to implement the CYGNO computing model, three major areas of development have been identified and prioritized: Cloud storage integration; support for interactive analysis; integration of batch processing. Furthermore, as detailed in sec. 6, additional developments for the online pre-processing are also foreseen and already planned, although with lower priorities.

5.1 Exploiting the INFN Cloud storage

The first objective has been to integrate the CYGNO Data Acquisition (DAQ) with the INFN Cloud object storage. The DAQ is based on the MIDAS framework [9] that takes care of acquiring camera images triggered by the PMTs signals when few photons (down to a single one) are detected in the TPC volume; auxiliary channels and high voltage are also monitored. Data and metadata are stored locally on disk and local SQL DB. Acquired data needs to be copied to the INFN Cloud Storage while the related metadata goes in the SQL database, where data and metrics are presented by means of Grafana interface. Calibration data are also produced, stored and used online and during the events reconstruction. In order to stage DAQ files to cloud we decided to use the object storage service offered by INFN Cloud providing S3 interfaces. It is implemented by a MinIO cluster, actually a Gateway for the Swift backend. As such, to stage data into the cloud storage, one needs to interface with MinIO. Minio is an open source distributed object storage server, providing

S3 storage functionality and a High Performance Object Storage API compatible with Amazon S3 cloud storage service. The latter actually represents a key feature for the CYGNO integration because it allows to use the Boto3 Python SDK for Amazon Web Services (AWS) as well to rely on MinIO Security Token Service (STS) for authentication (*AuthN*) and authorization (*AuthZ*) matters. In turn this means that on the one hand, the use of Boto3 grants us to operate on S3 buckets resources, creating, updating, and deleting files from Python scripts. On the other hand STS *AssumeRoleWithWebIdentity* API endpoint allows us to generate temporary access credentials using a JSON Web Token (JWT) returned from a configured OpenID Identity Provider such as INDIGO Identity and Access Management (IAM). The latter is the OpenID Connect Identity Provider adopted by INFN Cloud and has been integrated with MinIO as external Identity provider. Managing authorization rules to define who can read/write on buckets is a strong requirement for the experiment because acquired data must be properly secured. Finally the DAQ system relies on oidc-agent [10] software in order to locally manage the JWT.

5.2 Interactive data analysis

The ultimate goal is to provide CYGNO researchers with a "Software as a Service" like platform that facilitates the access to data for interactive analysis via a web interface. Since the goal is to deliver a multi-user platform, a JupyterHub based system has been put in place and, once again, the integration with INDIGO IAM has been made in order to manage the user *authN/Z*.The described system has been fully integrated within the TOSCA plus Ansible based system of INFN Cloud service portfolio (Figure 4).

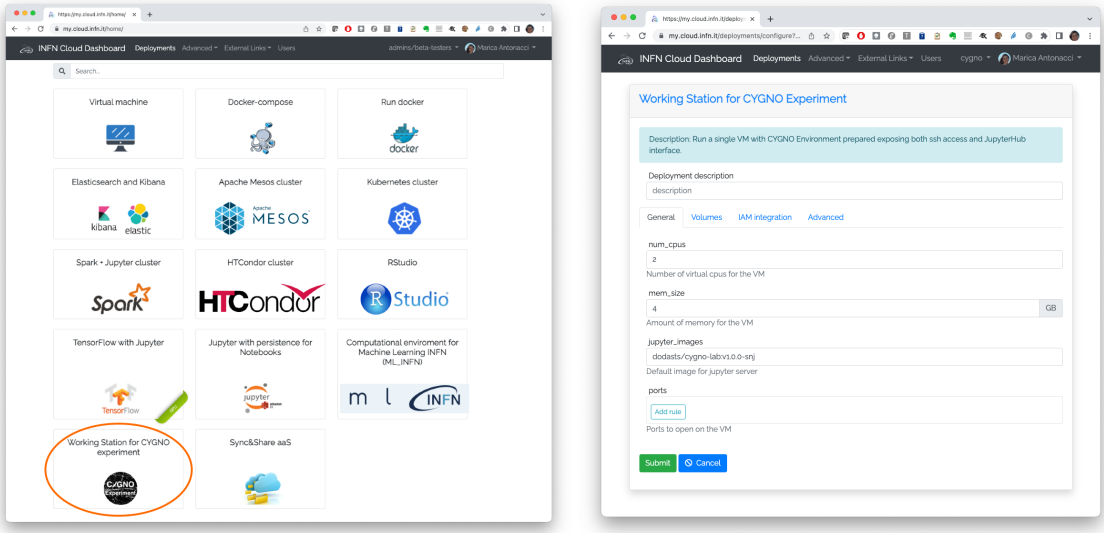

Figure 4: The INFN Cloud Dashboard. On the left: the service portfolio implemented so far. On the right: the configuration form that allows to customize the CYGNO compute environment to be deployed on Cloud.

Following the cloud native development model, since the beginning the runtime environment of the experiment (i.e. the required software packages and libraries) has been setup and managed via containers (Docker). A CYGNO golden image has been produced. Any user is allowed then to customize and personalize it based on specific needs. Technically those images are spawned

by JupyterHub anytime a user requires a JupyterLab instantiation. To this end the login form of JupyterHub supports the definition of the docker image to be spawned. A final remark is about the integration with cloud storage. A part of the S3 based access, the golden image is equipped with a tool based on rclone that is in charge of setting up an authenticated posix mount point that allows the user to access S3 via posix.

The implemented solution allows to access via a JupyterLab Web App to:

- ensure controlled access for CYGNO users to data and computing resources;
- develop and execute interactively Python/ROOT kernels algorithms;
- access data via POSIX and REST API to ensure local/remote and anonymous - open - access;
- provide environment for interactive data reconstruction, analysis and simulation (based on GEANT4/GARFIELD package);
- provide access to the experiment HTCondor batch queue for reconstruction, simulation and data analysis;
- provide data backup on TAPE;
- develop an orchestrator scale and provide the resources elastically on demand.

The CYGNO workflow - from DAQ to storage, analysis and reconstruction - has been developed and tested with the high background data acquired overground where the radioactivity is many order of magnitude larger than in the underground laboratory of LNGS.

Figure 5 shows an example of the CYGNO reconstruction algorithm working on an image (left) taken by the LIME prototype placed overground at the Laboratori Nazionali di Frascati (LNF) and its output (right), where it is possible to observe the found clusters.

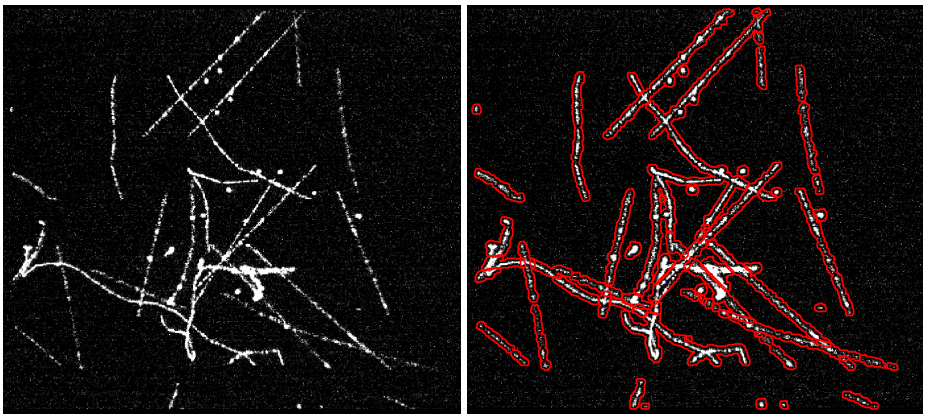

Figure 5: Example of an image taken overground by the LIME prototype (Left) and the clusters found by the collaboration reconstruction algorithm highlighted (Right).

5.3 On-Demand Batch System for CYGNO

The third development area is focused on enabling batch processing. A main requirement was to enable dynamic on demand capability to create and access batch resources. This has been realized by reusing an already available solution provided by INFN Cloud where a Kubernetes based service is used to deploy a complete HTCondor batch. The batch system is based on HTCondor 9.x. as this version grants support for token authentication (SCITOKENS method). In addition, being it completely cloudified, using custom containers and managing user tailored Workernodes is pretty easy. More important, the choice seems particularly suitable also because the underlying technology grants the possibility to easily scale up and down, a main motivation for CYGNO to move toward an elastic strategy for resource exploitation. The batch system has been fully integrated with the JupyterHub system described in sec 5.2 by meaning that a single JupyterLab interface allows to submit also batch jobs. The technical choice to move toward a container orchestration based platform goes in the direction of further extending the CYGNO cloud system with further services.

6. Future developments: Middle Ware Project

The Middle Ware (MW) is a project under development in order to evaluate data quality and produce a quasi online pre-reconstruction and analysis. The idea is to develop a framework, hosted in the INFN Cloud, capable of processing the data acquired by the detector and the sensors, delivery quality information about the runs, store information to the SQL DB and show the analyzed information with Dashboards.

Today a prototype is deployed on local resources and is under test, pushing reconstructed information and analysis on the SQL DB in Cloud. The SQL DB is accessed by a Grafana server, also placed in the INFN Cloud, that generates useful real-time information about the analyzed data, as illustrated in Figure 6.

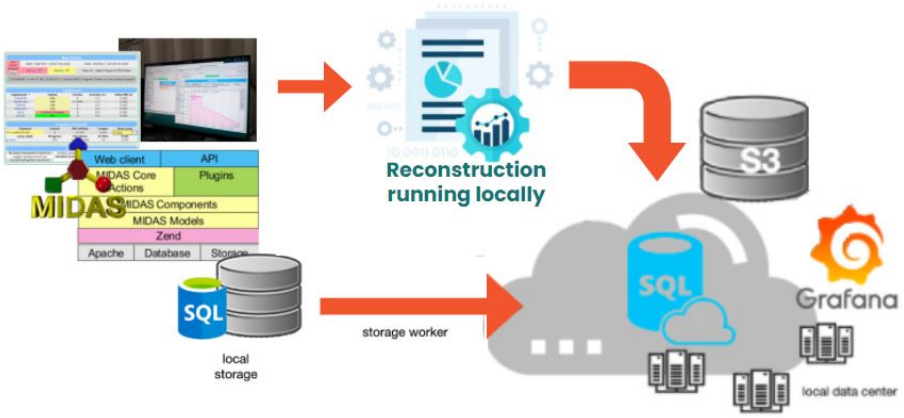

Figure 6: Nowadays Data Flow - schematic view from detector to cloud.

For the next step, the goal is to have the whole data pipeline hosted in the INFN Cloud, in a way that it is possible to scale the computational/storage needs when it will be necessary. One of the foreseen possibilities is to place the Apache Kafka Software [11] as a data worker to handle the

real-time data and feed the reconstruction algorithm managed by the Apache Spark software [12], both software hosted in the cloud, as illustrated in the Figure 7.

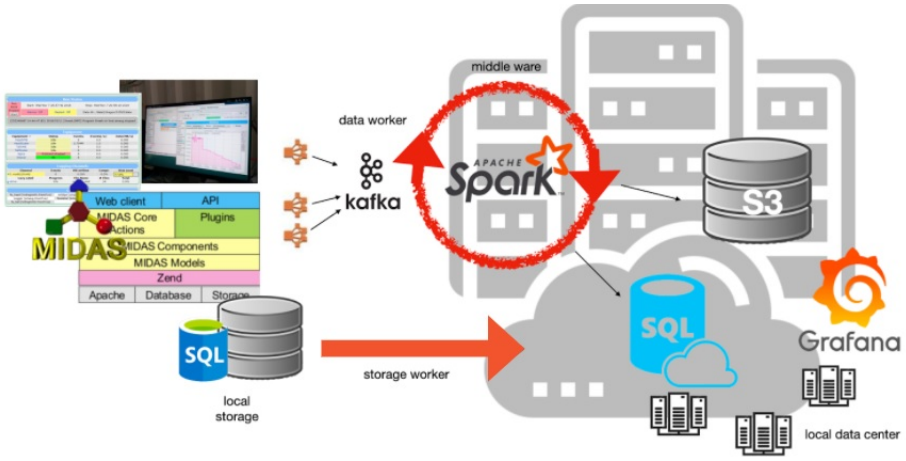

Figure 7: Future Data Flow - schematic view from detector to cloud.

7. Conclusions

In conclusion, INFN Cloud through its Platform as a Service (PaaS) layer allows the experiments to access resources as a Software as a Service (SaaS), and CYGNO is the beta-tester of this system. This implementation did not show any conceptual or practical limits that the cloud architecture has not been able to solve.

The cloud, beyond specific experiments, seems to be the best way to efficiently manage the available resources, in particular for small and medium-sized experiments, now more and more frequent not only in the world of astroparticles.

Acknowledgement

We thank M. Punturo for the useful discussion and hints. This CYGNO/INTIUM project has received fundings under the European Union's Horizon 2020 research and innovation programme from the European Research Council (ERC) grant agreement No 818744.

References

[1] S. Mohanty, *Astroparticle Physics and Cosmology: Perspectives in the Multimessenger Era*, vol. 975. Springer International Publishing, 2020, 10.1007/978-3-030-56201-4".

[2] F. D. Amaro et al., *The CYGNO Experiment*, *Instruments* **6** (2022) 6 [2202.05480].

[3] L. M. S. Margato, F. A. F. Fraga, S. T. G. Fetal, M. M. F. R. Fraga, E. F. S. Balau, A. Blanco et al., *Performance of an optical readout GEM-based TPC*, *Nucl. Instrum. Meth.* **A535** (2004) 231.

[4] S. E. Vahsen, C. A. J. O'Hare, W. A. Lynch, N. J. C. Spooner and E. Baracchini, *CYGNUS: Feasibility of a nuclear recoil observatory with directional sensitivity to dark matter and neutrinos*, *arXiv:2008.12587* (2020) .

[5] I. A. Costa, E. Baracchini, F. Bellini, L. Benussi, S. Bianco, M. Caponero et al., *Performance of optically readout GEM-based TPC with a ^{55}Fe source*, *Journal of Instrumentation* **14** (2019) P07011.

[6] E. Baracchini, L. Benussi, S. Bianco, C. Capoccia, M. Caponero, G. Cavoto et al., *Identification of low energy nuclear recoils in a gas time projection chamber with optical readout*, *Measurement Science and Technology* **32** (2020) 025902.

[7] E. Baracchini, L. Benussi, S. Bianco, C. Capoccia, M. Caponero, G. Cavoto et al., *A density-based clustering algorithm for the CYGNO data analysis*, *Journal of Instrumentation* **15** (2020) T12003.

[8] A. Abc, *Infn cloud project (we will add it soon)*, *Journal* **1** (2016) 145.

[9] PSI and TRIUMF, "MIDAS modern data acquisition page." https://daq00.triumf.ca/MidasWiki/index.php/Main_Page.

[10] A. Abc, *We will add it soon*, *Journal* **1** (2016) 145.

[11] M. J. Sax, *Apache kafka*, in *Encyclopedia of Big Data Technologies*, S. Sakr and A. Zomaya, eds., (Cham), pp. 1–8, Springer International Publishing, (2018), DOI.

[12] S. Salloum, R. Dautov, X. Chen, P. X. Peng and J. Z. Huang, *Big data analytics on apache spark*, *International Journal of Data Science and Analytics* **1** (2016) 145.