EPJ Web of Conferences **295**, 07013 (2024) https://doi.org/10.1051/epjconf/202429507013
CHEP 2023

Data handling of CYGNO experiment using INFN-Cloud solution

F.D. Amaro[1], *M.* Antonacci[17], *R.* Antonietti[2,3], *E.* Baracchini[4,5], *L.* Benussi[6], *S.* Bianco[6], *F.* Borra[7,8], *A.* Calanca[6], *C.* Capoccia[6], *M.* Caponero[6,9], *D.S.* Cardoso[10], *G.* Cavoto[7,8], *D.* Ciangottini[16], *I.A.* Costa[6], *G.* D'Imperio[8], *E.* Dané[6], *G.* Dho[4,5], *F.* Di Giambattista[4,5], *E.* Di Marco[8], *C.* Duma[15], *F.* Iacoangeli[8], *H.P.* Lima Júnior[10], *E.* Kemp[11,5], *G.S.P.* Lopes[12], *G.* Maccarrone[6], *R.D.P.* Mano[1], *R.R.* Marcelo Gregorio[13], *D.J.G.* Marques[4,5], *G.* Mazzitelli[6,*], *A.G.* McLean[13], *P.* Meloni[2,3], *A.* Messina[7,8], *C.M.B.* Monteiro[1], *R.A.* Nobrega[12], *I.F.* Pains[12], *E.* Paoletti[6], *L.* Passamonti[6], *C.* Pellegrino[15], *F.* Petrucci[2,3], *S.* Piacentini[7,8], *D.* Piccolo[6], *D.* Pierluigi[6], *D.* Pinci[8], *A.* Prajapati[4,5], *F.* Renga[8], *R.J.d.C.* Roque[1], *F.* Rosatelli[6], *A.* Russo[6], *J.M.F.* dos Santos[1], *G.* Saviano[14,6], *D.* Spiga[16], *N.J.C.* Spooner[13], *S.* Stalio[4], *R.* Tesauro[6], *S.* Tomassini[6], *S.* Torelli[4,5],

[1]LIBPhys, Department of Physics, University of Coimbra, 3004-516 Coimbra, Portugal
[2]Dipartimento di Matematica e Fisica, Università Roma TRE, 00146, Roma, Italy
[3]Istituto Nazionale di Fisica Nucleare, Sezione di Roma Tre, 00146, Rome, Italy
[4]Gran Sasso Science Institute, 67100, L'Aquila, Italy
[5]Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali del Gran Sasso, 67100, Assergi, Italy
[6]Istituto Nazionale di Fisica Nucleare, Laboratori Nazionali di Frascati, 00044, Frascati, Italy
[7]Dipartimento di Fisica, Università La Sapienza di Roma, 00185, Roma, Italy
[8]Istituto Nazionale di Fisica Nucleare, Sezione di Roma, 00185, Rome, Italy
[9]ENEA Centro Ricerche Frascati, 00044, Frascati, Italy
[10]Centro Brasileiro de Pesquisas Físicas, Rio de Janeiro 22290-180, RJ, Brazil
[11]Universidade Estadual de Campinas - UNICAMP, Campinas 13083-859, SP, Brazil
[12]Universidade Federal de Juiz de Fora, Faculdade de Engenharia, 36036-900, Juiz de Fora, MG, Brasil
[13]Department of Physics and Astronomy, University of Sheffield, Sheffield, S3 7RH, UK
[14]Dipartimento di Ingegneria Chimica, Materiali e Ambiente, Sapienza Università di Roma, 00185, Roma, Italy
[15]INFN-CNAF, Viale Carlo Berti Pichat 6/2, 40127 Bologna, Italy
[16]Istituto Nazionale di Fisica Nucleare, Sezione di Perugia, 06123, Perugia, Italy
[17]Istituto Nazionale di Fisica Nucleare, Sezione di Bari, 70125, Bari, Italy

Abstract. The INFN Cloud project was launched at the beginning of 2020, aiming to build a distributed Cloud infrastructure and provide advanced services for the INFN scientific communities. A Platform as a Service (PaaS) was created inside INFN Cloud that allows the experiments to develop and access resources as a Software as a Service (SaaS), and CYGNO is the beta-tester of this system. The aim of the CYGNO experiment is to realize a large gaseous Time Projection Chamber based on the optical readout of the photons produced in the avalanche multiplication of ionization electrons in a GEM stack. To this extent, CYGNO exploits the progress in commercial scientific Active Pixel Sensors based on Scientific CMOS for Dark Matter search and Solar Neutrino studies. CYGNO, like many other astroparticle experiments,

*e-mail: giovanni.mazzitelli@lnf.infn.it

requires a computing model to acquire, store, simulate and analyze data typically far from High Energy Physics (HEP) experiments. Indeed, astroparticle experiments are typically characterized by being less demanding of computing resources with respect to HEP ones but have to deal with unique and unrepeatable data, sometimes collected in extreme conditions, with extensive use of templates and montecarlo, and are often re-calibrated and reconstructed many times for a given data set. Moreover, the varieties and the scale of computing models and requirements are extremely large. In this scenario, the Cloud infrastructure with standardized and optimized services offered to the scientific community could be a useful solution able to match the requirements of many small/medium size experiments. In this work, we will present the CYGNO computing model based on the INFN cloud infrastructure where the experiment software, easily extendible to similar experiments to similar applications on other similar experiments, provides tools as a service to store, archive, analyze, and simulate data.

1 Introduction

The road-map on searching dark matter and investigate the nature of the neutrinos is clear. From one side it is based on the construction and improvements of large worldwide experiments like DARKSIDE, PANDA, LUX, XENON etc. [1], which are trying to detect anomalous events in the nuclear recoils while from the other side the efforts are focused on increasing volumes and performances of solid detectors able to investigate dark matter in the very low mass sector. However, these detectors could have some limitation, as they are reaching the neutrino floor, where it's not easy to discriminate dark matter candidates from neutrino-induced nuclear recoils, and at the same time they will never been able to make a "dark matter astronomy" because they are not able to detect the signal direction. A similar situation concerns solar neutrinos, where large liquid detectors in experiments like Hyper-Kamiokande, JUNO, etc. [2] could have limited energy threshold for the detection or are not able to fully identify the source direction.

In this scenario, He based gaseous Time Projection Chambers (TPCs) could be a candidate for future improvements for such signals detection and dark matter and solar neutrinos astronomy. For this reason a world wide proto collaboration, CYGNUS, is studying and developing gaseous TPCs wherein the CYGNO [3] project at INFN Laboratori Nazionali del Gran Sasso (LNGS) is one of the most active and head. CYGNO is exploiting the progress in commercial scientific Active Pixel Sensors (APS) based on CMOS technology to realize a large gaseous Time Projection Chamber (TPC) [5]. Charges produced by ionization due to particles interaction in the gas release electrons (and ions) that drift in the 50kV electric filed of the TPC reaching a triple GEMs stage of amplifications[4].

The GEMs produce an avalanche of electrons and photons in the visible spectra that can be exploited to identify particle tracks features such as path, direction, energy lost, etc.

The CMOS technology, although slow for many HEP applications, allows to equip millions equivalent readout channels with single photons noise (comparable with to 0.4 keV energy threshold and 150 micron of spatial resolution) with a reasonable similar cost per channel with respect to electronic ones. This feature is promising for the future when TPC detector with high granularity will need to scale up their dimensions.

CYGNO group is ending the R&D phase with a prototype, called LIME [7], installed at LNGS, and has been funded by INFN and EU with the INITUM grant, to realize a demonstrator whose installation is scheduled for 2025 to prove that the technology is mature for the construction of O($30m^3$) detector able to perform physics measurements (see fig. 1).

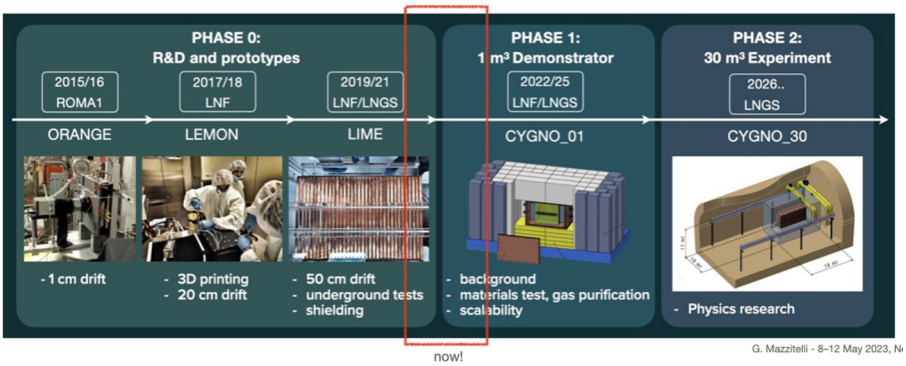

Figure 1. CYGNO project road map. To date, LIME is finishing the R&D by testing data taking, calibration and reconstruction and also analysis algorithms, comparing data and montecarlo full simulation, validating ancillary system like gas system, DAQ and computing infrastructure. The preparation of the site to host CYGNO04 [6] has started and in 2025 the final installation of the demonstrator is expected.

LIME prototype is also the device employed for testing DAQ, computing model and the infrastructure for CYGNO04 demonstrator and future larger detectors.

2 Requirements

LIME is equipped by a single camera, 2304×2304 resolution, and 4 PMTs, symmetrically placed around the camera to detect the time shape longitudinal evolution. The typical DAQ event size is 12MB and the data rate is determined by the ability to shield the beta/gamma background, that in LIME, with its 50 liters of sensitive volume, is today about 1Hz. The surviving beta/gamma background has to be discriminated by the reconstruction and analysis software[8, 9], in order to identify possible nuclear recoils. The reconstruction software applies a set of thresholds, filters and clusterization algorithms producing subsets of data each containing single track events information. Because of thresholds and filters, also raw data must be saved to be possibly reprocessed. The today LIME throughput is about 400Gb/day. A larger demonstrator like, CYGNO04, is expected to have an event size four times more, and a real detector able to achieve physics goals, about a hundred times more with respect to LIME throughput. The event rate will depend on the ability to reduce background by means of optimal, low radioactivity, construction materials and the layers of shielding materials to stop beta/gamma components and residual natural neutrons. This situation requires to design and test a computing model and the hosting computing infrastructure with features and capable to digest and elaborate this challenging scenario.

Moreover, a large amount of computing resources are needed for simulation of particle interactions with the gas in the detector, their transport in the drift field, the evaluation of the light produced and finally the digitization of the signal in order to compare the simulation results with the images collected in the data.

3 Computing model and infrastructure

The CYGNO project, namely LIME today, is hosted in the underground laboratory of LNGS where it is recommended to have only the minimum setup necessary to collect data (DAQ) equipped with a local buffer to ensure that no data is lost in case of network or back-end

faults. Many experiments up to now decided to host their back-end computing infrastructure in the overground computing center of LNGS, to ensure proximity of back-end apparatus, powers and computing rooms suitable for hosting the necessary servers. This requires the experiments and the laboratory efforts to setup, put in operation and maintain all the infrastructure. Those efforts could be an issue for a small/medium experiment typical of astro-particle physics area (see conclusion in chapter 5).

In 2020 the INFN-Cloud [10] project started, offering many services at PaaS/SaaS level, optimal to host the CYGNO computing model, ensuring the characteristics of scalability, safety, reliability etc. This infrastructure allows the collaboration, together with the INFN-Cloud, to develop and integrate a set of tools for data management, analysis and simulation available at user level and accessible and exploitable to all the CYGNO international collaborators. In particular this has been developed through a set of inter-operated services, initially based on "Dynamic On Demand Analysis Service (DODAS)" [11] project.

3.1 Architecture of the system

The main pillars of the architecture have been defined taking into account the requirements gathered from CYGNO researchers. The identified needs span from the support interfaces for interactive analysis to on demand batch systems passing through data management and monitoring solutions. In details the main requirements are summarized below:

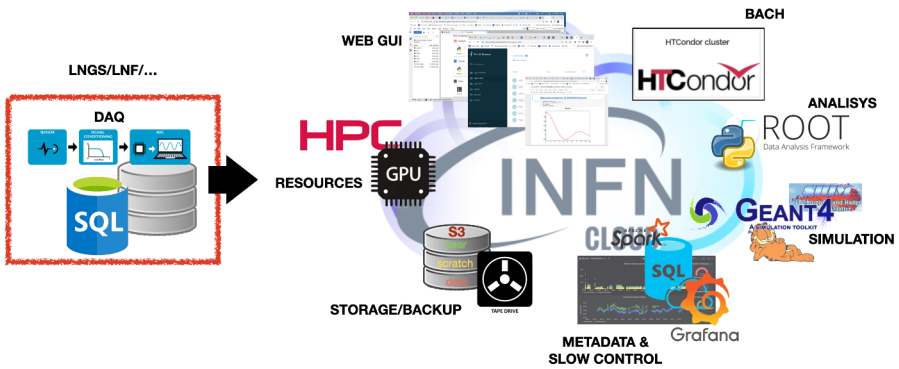

Figure 2. schematic view of back-end services developed on INFN-Cloud

- user need to run notebooks as well as to access consoles services for software development purposes and interactive data analysis performed via python ecosystem and ROOT.
- working environment should support data access via POSIX (FUSE simulated).
- Persistent S3 storage is also required to support middle ware full reconstruction. The Data Transformation Services (DTS) fully reconstructing and calibrating data producing ROOT files for analyses, stored in the S3 experiment area.
- A system to fully automate the continuous backup system to archive raw data on tape library.
- On demand batch system are needed in order to process data reconstruction and simulation. The system should be configurable within the experiments requirements (libraries and dependencies)
- a set of sql/nosql database where experiments meta data are collected and historicized.

- GRAFANA based interface where experiment meta data are presented.
- middle ware services: a set of services based on DAQ steamed data: remote console, fast reconstruction and quality monitor (see section 4).

The technology choices have been made in order to better fit into the INFN-Cloud service model based on the paradigm that researchers must be allowed to exploit "free" and open services to manage workflows, build pipelines, data processing and analysis and, of course, to share/to reuse technical solutions. This is what allow researchers to focus on science. In order to achieve all this two technical drivers are identified:

- to enable users to create and provision infrastructure deployments, automatically and repeatedly, with almost zero effort.
- to implement the Infrastructure as Code paradigm based on declarative approach: allows to describe "What" instead of "How"

The only user handled part must be the runtime computational environment. Technologically wise this has been achieved promoting and supporting container-based solutions at all level of the architecture.

Concretely CYGNO developed a single entry-point based on JupyterHub fully integrated with INDIGO-IAM OAuth2 authorization service in order to support JWT based authentication and authorization. Containerization to allow user to customize their run-time environment, the Jupyterlab environment. This system natively support the integration with on-demand batch system implemented using HTCondor. From a technical perspective HTCondor batch system on demand are deployed as k8s application. Worker-nodes are also containerized and thus users can easily customize the run-time environment as they prefer. Currently the interactive and batch environments are configured in the same way, with all the requirements for analysis (generic and experiment libraries) and simulation software (e.g. GEANT, GARFIELD), in order to ensure fast and easy setup via notebook before move to full analysis and simulation on the batch system (see fig. 2).

Finally, the interactive environment has been developed in order to automatically mount S3 MinIO buckets, using wrapper around RClone provided by INFN-Cloud. The wrapper has been named sts-wire and it guarantees a proper handling of the authentication and authorization that relies on INDIGO-IAM JWT. At batch level the interaction with S3 is performed via pre-sgined url. The next section details the CYGNO Middle Ware project that mainly cares about the computing aspects of the data acquisition system.

4 CYGNO Middle Ware project

Data produced by the CYGNO DAQ based on the MIDAS framework [12], are streamed, event by event, and collected by means of a *kafka* process into the INFN-Cloud where are processed online to monitor data quality and stability. Also the MIDAS DAQ Online DataBase (ODB) is streamed in cloud, where GRAFANA show in real time runs information and sub samples of raw data (fig. 3). MIDAS ODB is also historicized by means of nosql db with a sub-sample rate (10 minutes).

The streaming service (fig. 4) is configured to be fast but low priority - data lossy - and any event caught is continuously sent to a fast pipeline of analysis and the output is dumped in nosql db and presented in the GRAFANA experiment dashboards.

Moreover, raw files are copied when 400 events are collected ($\sim$ 1.5Gb) into the S3 experiment storage on Cloud backbone. Runs meta-data are also update by DAQ and stored locally in a *MariaDB* mirrored in the equivalent INFN-Cloud services (see fig. 5). A process, *tape-r*, continuously monitor the presence of data in the remote storage S3 and automatically

Figure 3. The CYGNO remote console available for user following the run from remote - without the need to access the DAQ, restricted to experts - where the following are available: device setting and readout, sample of raw images collected and PMT waveforms, typically updated every 5 seconds.

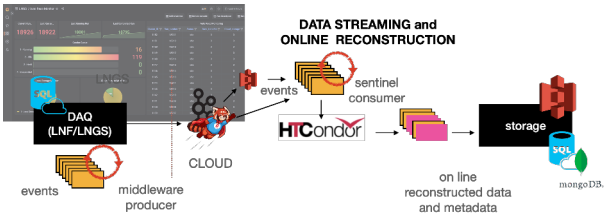

Figure 4. CYGNO on line reconstruction flow

backup it on TAPE. A few minutes after run is closed data are available locally, remotely - ready for the analysis - and backed up on TAPE.

A second pipeline of restriction, run in parallel to the data lossy one, is based on streamed data: as soon as the data, packed in runs of 400 events, are available on S3 the *sentinel* process submit it to the condor queue for the full reconstruction, producing official ROOT files (DTS) for analysis available for the collaboration. About half an hour after the run is closed, DTS are available for analyses. The same *sentinel* process can be exploited to reprocess data if needed. Furthermore, a process - *analyzer* - exploits the DTS files producing the most common histograms and history to double check the data quality and reconstruction procedure.

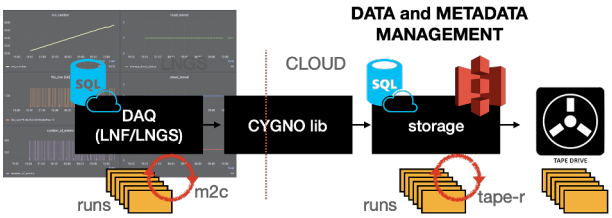

Figure 5. CYGNO data management system

All the processes are developed and configured in dedicated docker containers running on Virtual Machines (VMs) in the INFN-Cloud interconnected by means of secure tunnels, in order to exchange information between both LNGS and the Cloud, as well as between VMs in the Cloud.

The local storage at LNGS (10TB RAID) allows to have about one month data on local buffer, well above the minimum requirement to ensure data preservation in case of network issues, transfer error etc. The experiment storage (remote) consists of 100TB on S3, able to store a typical run data set of some months and retrieve from tape an equivalent set of data for different reconstructions. About 10% is dedicated to simulations and analysis output.

200TB are available on TAPE for final staging of the data. Dedicated GRAFANA dashboards monitor data flow, properties and alerts of failures.

Finally, the future direction is to develop a Rucio[13] based system to manage the CYGNO data. Rucio is open-source software licensed under Apache v2.0, and makes use of established open-source toolchains. Among its main features it manages location-aware data in a heterogeneous distributed environment, including creation, location, transfer, deletion, and annotation. Declarative orchestration of dataflows with both low-level and high-level policies is its main mode of operation.

5 Conclusion

CYGNO is designing and testing a set of services based on the INFN-Cloud infrastructure and its future evolution supported by the ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing [14]. The developed computing model and the INFN-Cloud infrastructure appear to answer to the technical requirements of CYGNO experiment. Moreover, the choice of the computing infrastructure ensures the possibility of scaling, flexibility, reliability, efficiency required by an R&D moving from a demonstrator to a large detector with strong demanding requirements.

The services developed for the CYGNO computing model, can be generalized and scaled for many other applications in the astro-particle experiments. Indeed, CYGNO is a typical small/medium experiment of astro-particle community. This community is characterized mainly by experiments with a relative low rate and large event data size. However, the data management is not less challenging than any other experiment with high throughput (fig. 6). Nevertheless, most of these experiments have common requirements that can be reached ex-

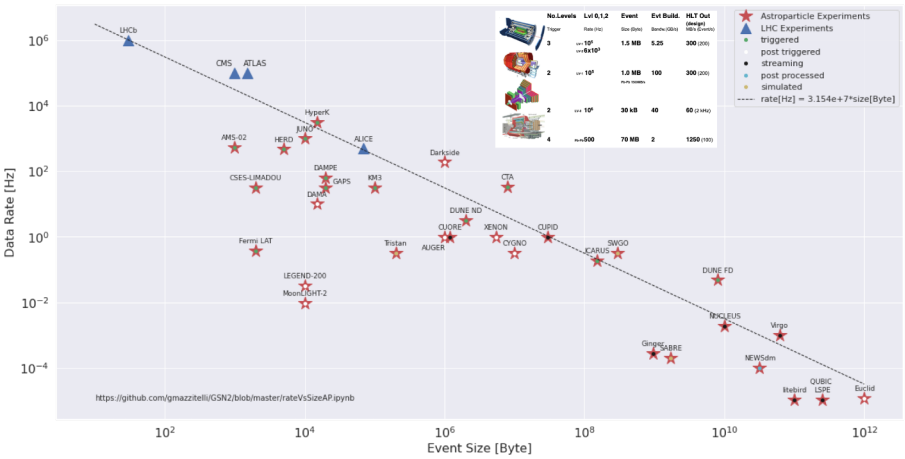

Figure 6. Results of the survey on the Italian astro-particle community about the characteristics of their computing model. The experiments are characterized by having a different throughput with respect to typical HEP experiments. It is possible to see that even if they are not HEP experiments, they follow a scaling law that underline how are demanding in the overall process.

ploiting the development of generalizable and scalable services a SaaS level on a powerful cloud infrastructure, as the INFN one.

6 Acknowledgement

This work is partially supported by ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, funded by European Union – NextGenerationEU and by ERC-INITIUM-818744

References

[1] J. Billard, *et al.* Rept. Prog. Phys. **85**, no.5, 056201 (2022) doi:10.1088/1361-6633/ac5754 [arXiv:2104.07634 [hep-ex]].

[2] Pallavicini, M. Solar neutrinos: experimental review and prospectives. *Journal Of Physics: Conference Series*. **598**, 012007 (2015,3), https://dx.doi.org/10.1088/1742-6596/598/1/012007

[3] Amaro, F., *et al.* The CYGNO Experiment. *Instruments*. **6** (2022), https://doi.org/10.3390/instruments6010006

[4] Sauli, F. GEM: A new concept for electron amplification in gas detectors. *Nuclear Instruments And Methods In Physics Research Section A: Accelerators, Spectrometers, Detectors And Associated Equipment*. **386**, 531-534 (1997), https://doi.org/10.1016/S0168-9002(96)01172-2

[5] Antochi, V., *et al.* A GEM-based optically readout time projection chamber for charged particle tracking. *ArXiv*. (2020)

[6] Mazzitelli, G., *et al.* Technical Design Report - TDR CYGNO-04/INITIUM. (2023,2), https://doi.org/10.15161/oar.it/76967

[7] Mazzitelli, G., *et al.* D. 50 litres TPC with sCMOS-based optical readout for the CYGNO project. *Nuclear Instruments And Methods In Physics Research Section A: Accelerators, Spectrometers, Detectors And Associated Equipment*. **1045** pp. 167584 (2023), https://doi.org/10.1016/j.nima.2022.167584

[8] Baracchini, E.,, *et al.* A density-based clustering algorithm for the CYGNO data analysis. *Journal Of Instrumentation*. **15**, T12003 (2020,12), https://dx.doi.org/10.1088/1748-0221/15/12/T12003

[9] Amaro, F., *et al.* iDBSCAN to detect cosmic-ray tracks for the CYGNO experiment. *Measurement Science And Technology*. **34**, 125024 (2023,9), https://dx.doi.org/10.1088/1361-6501/acf402

[10] Amaro, F., *et al.* Exploiting INFN-Cloud to implement a Cloud solution to support the CYGNO computing model. https://doi.org/10.22323/1.415.0021

[11] INDIGO-DataCloud, Dynamic On Demand Analysis Service (DODAS), https://web.infn.it/dodas/

[12] PSI & TRIUMF, MIDAS modern data acquisition, https://daq00.triumf.ca/MidasWiki/

[13] Barisists M et al. Rucio: Scientific Data Management, https://link.springer.com/article/10.1007/s41781-019-0026-3

[14] ICSC – Centro Nazionale di Ricerca in High Performance Computing, Big Data and Quantum Computing, https://www.supercomputing-icsc.it/