Astronomy ESFRI & Research Infrastructure Cluster

ASTERICS - H2020 - 653477

Resources and Requirements Analysis

ASTERICS GA DELIVERABLE: D3.5

Document identifier:	ASTERICS-D3.5.docx
Date:	**31 October 2016**
Work Package:	**WP3 Obelics**
Lead Partner:	**INAF / ASTRON**
Document Status:	**Final**
Dissemination level:	**Public**
Document Link:	www.asterics2020.eu/documents/ ASTERICS-D3.5.pdf

Abstract

A Resources and Requirements Analysis is made of several astronomy and astroparticle ESFRI projects and research infrastructures linked in the ASTERICS project. All these facilities expect to deal with unprecedented amount of generated and/or reconstructed data during their operation. The major challenges concern the storage and the transfer of those data, their access through complex but efficient databases and their analysis on distributed computational resources, in a reasonable amount of time.

I. COPYRIGHT NOTICE

II. DELIVERY SLIP

	Name	Partner/WP	Date
From	T. Vuillaume	CNRS-LAPP	
Author(s)	T. Vuillaume T.J Dijkema	CNRS-LAPP ASTRON	
Reviewed by	R. van der Meer, G. Cimo	ASTRON	
Approved by	AMST		31/10/2016

III. DOCUMENT LOG

Issue	Date	Comment	Author/Partner
1	21-06-2016	Initial version CTA	Thomas Vuillaume, LAPP
2	13-07-2016	Updated with LOFAR/SKA	Tammo Jan Dijkema, ASTRON
3	28-07-2016	Updated version CTA	Thomas Vuillaume, LAPP

4	30-08-2016	Updated version KM3NET	Kay Graf, FAU
5	15-09-2016	Updated SKA	Peter Hague
6	19-09-2016	LSST part from N. Chottard	T. Vuillaume
7	23-10-2016	Review, format	T. Vuillaume
6	28-10-2016	Updated from online version	Tammo Jan Dijkema, ASTRON

IV. APPLICATON AREA

This document is a formal deliverable for the GA of the project, applicable to all members of the ASTERICS project, beneficiaries and third parties, as well as its collaborating projects.

V. TERMINOLOGY

ASTERICS	Astronomy ESFRI & Research Infrastructure Cluster
CTA	Cherenkov Telescope Array
E-ELT	European Extremely Large Telescope
ESFRI	European Strategy Forum on Research Infrastructures
KM3NeT	Cubic Kilometre Neutrino Telescope
LOFAR	The Low Frequency Array
LSST	The Large Synoptic Survey Telescope
OBELICS	Observatory E-environments LInked by common ChallengeS
SKA	The Square Kilometre Array
MEGA-	prefix for one million (10^6)

GIGA-	prefix for one billion (10^9)
TERA-	prefix for 1000 billion (10^{12})
PETA-	prefix for one million billion (10^{15})
EXA-	prefix for one billion billion (10^{18})

VI. PROJECT SUMMARY

ASTERICS (Astronomy ESFRI & Research Infrastructure Cluster) aims to address the cross-cutting synergies and common challenges shared by the various Astronomy ESFRI facilities (SKA, CTA, KM3Net & E-ELT). It brings together for the first time, the astronomy, astrophysics and particle astrophysics communities, in addition to other related research infrastructures. The major objectives of ASTERICS are to support and accelerate the implementation of the ESFRI telescopes, to enhance their performance beyond the current state-of-the-art, and to see them interoperate as an integrated, multi-wavelength and multi-messenger facility. An important focal point is the management, processing and scientific exploitation of the huge datasets the ESFRI facilities will generate. ASTERICS will seek solutions to these problems outside of the traditional channels by directly engaging and collaborating with industry and specialised SMEs. The various ESFRI pathfinders and precursors will present the perfect proving ground for new methodologies and prototype systems. In addition, ASTERICS will enable astronomers from across the member states to have broad access to the reduced data products of the ESFRI telescopes via a seamless interface to the Virtual Observatory framework. This will massively increase the scientific impact of the telescopes, and greatly encourage use (and re-use) of the data in new and novel ways, typically not foreseen in the original proposals. By demonstrating cross-facility synchronicity, and by harmonising various policy aspects, ASTERICS will realise a distributed and interoperable approach that ushers in a new multi-messenger era for astronomy. Through an active dissemination programme, including direct engagement with all relevant stakeholders, and via the development of citizen scientist mass participation experiments, ASTERICS has the ambition to be a flagship for the scientific, industrial and societal impact ESFRI projects can deliver.

VII. EXECUTIVE SUMMARY

A Resources and Requirements Analysis is made of several astronomy and astroparticle ESFRI projects and research infrastructures linked in the ASTERICS project. All these facilities expect to deal with unprecedented amount of generated and/or reconstructed data during their operation. The major challenges concern the storage and the transfer of those data, their access through complex but efficient databases and their analysis on distributed computational resources, in a reasonable amount of time.
It is conceptually clear that all the facilities can benefit from synergy in how they handle the large data volumes and databases. This study identifies further which areas are most interesting to start the development of common solutions, standards and tools.
As it is the first time that these facilities and the fields they were developed in work together, this study started with the description of the expected data structures from each facility. When commonalities are found a more detailed investigation of the opportunities can be excuted. The partners were asked to describe the foreseen data handling and storage in their own way.
There are many differences, but the overview of the provided numbers in the table at the end shows commonalities as well. the results serve as starting point for further detailed comparison and for collaborative developments.
The report will make clear that the projects under investigation here are in different states of development and maturity of deciding on data structure, which makes it difficult to find clear synergies and commonalities. By describing the projects data structure, the projects are already learning from each other.

Table of Contents

Table of Contents

1. Introduction

The ASTERICS projects aims at bringing together several astronomy and astroparticle ESFRI projects and research infrastructures. For the Resources and Requirements Analysis in this document several of these are compared. These are the Square Kilometer Array (SKA) and LOFAR as a precursor, the Cerenkov Telescope Array (CTA), KM3NeT, the EUCLID mission and the Large Synoptic Survey Telescope (LSST). All these projects have in common to deal with unprecedented amount of generated and/or reconstructed data. The major challenges concern the storage and the transfer of those data, their access through complex but efficient databases and their analysis on distributed computational resources, in a reasonable amount of time.

While it is conceptually clear that all the projects can benefit from synergy in how they handle the large data volumes and databases, it was at the beginning of the project not clear where this synergy can be found and how combined progress can be achieved. The first part of Task 3.3, Data Systems Integration (D-INT), is to compare the participating projects and find out where synergy in data handling and storage can be found. For this purpose, we asked each of the partners to contribute a section to this report on resource requirements.

The kind of data at hand in the different projects, and thus the way they are managed, can be very diverse. This document describes the data model of each instrument and focus on the resources required to manage the large data volumes, from their generation, to their archive and their analysis. It will serve as a comparison basis between the ESFRI projects to better understand their common ground and differences in their data models. This report does not pretend to be exhaustive or to cover all the aspects of the ESFRI projects data models but rather to give an overview and allow intelligible comparisons.

The report will make clear that the projects under investigation here are in different states of development and maturity of deciding on data structure, which makes it difficult to find clear synergies and commonalities. On the other hand, the comparison table with key numbers on the different instruments at the end of this report shows that we do have a lot in common, as has also been found out in project meetings. The Task 3.3 (D-INT) will continue to investigate commonalities between the projects and determine more detailed information on these commonalities in order to start developing tools and standards for these areas.

On asset already visible from this work is that projects in their early phase of defining their data structure can clearly learn from the more mature projects.

2. CTA

2.1. Introduction

The Cerenkov Telescope Array (CTA) is an imaging atmospheric Cerenkov Observatory for gamma-ray astronomy between 20 GeV (electronVolt) and 100 TeV whose construction is planned to start in early 2017.

2.2. Data Model

The CTA data flow transforms raw data recorded by the telescopes and consisting of *events* into high level scientific data product (light curves, sky maps and spectra) through different steps of the pipeline. The event data from the Cherenkov cameras is complemented by calibration and technical data which do not always follow the "triggered" event pattern, but are frequently acquired at equally spaced time intervals.

2.2.1. Data levels

The steps of the pipeline lead to five levels of data, each one representing a level of analysis, from Data Level 0 to Data Level 5 (see below). Different data types belonging to these levels and enabling the reconstruction of recorded events have been identified:

- event data (EVT) from the telescope cameras
- calibration data (CAL) coming from the telescope
- technical data (TECH) such as
    - engineering data (ENG) recorded by the telescopes (tracking, pointing, high voltage, temperature sensors, etc)
    - auxiliary (AUX) data related to the auxiliary systems with information related to sub-arrays or the full CTA array, including calibration data (i.e. weather, LIDAR profiles, etc.)
- Monte Carlo (MC) event data (simulated EVT data with additional information about the simulated particle characteristics) and instrumental response functions (IRF).

The data production is comprised of a series of processing steps that transform (reduce) archived raw data DL0 (Data Level 0) to calibrated camera data (DL1), then to reconstructed shower parameters such as energy, direction, and particle ID (DL2), and finally to high-level observatory products comprised of selected gamma-like events, instrument response tables, and housekeeping data (DL3). DL3 data will have a total volume of about 2% of the DL0 data volume and guaranteed access will be provided in the CTA archive to basic users. Science tools, to be provided by the observatory, will then be used either automatically or by users to produce DL4 (e.g. spectra, sky-maps). Finally, (DL5) legacy observatory data, such as CTA survey sky maps or the CTA source catalog will be produced.

2.2.2. Archives

It has been estimated that in the final CTA configuration, nearly 4PB/yr of DL0 data will be generated. However, after processing and when all replicas and versions are considered, the CTA archive should handle around 25 PB/yr of data. These data will be transferred and/or generated by four off-site data centers $(DC)_i$. The CTA archive is structured in different parts among which are the storage and the management systems. The storage system manages the physical data repositories according to specific data types and data levels and the management system includes all the products related to the archive databases and browsing systems.

Archive architecture will follow a division related to its content: Raw Data Archive, Calibration Data Archive, Engineering Data Archive, MC Data Archive and Science Data Archive. Of particular interest is the Science Data Archive, which stores the DL3 data that are the primary inputs of the CTA science tools. Archive users as well as Guest observers can have access to this archive according to CTA Data Delivery Policy. Official CTA science products like spectra, sky map and light curve (DL4) as well as CTA survey and catalogs (DL5) are stored in the High Level Multi-Frequency Data Archive (VO compliant).

2.2.3. Database management system

CTA DB architecture and systems have yet to be determined and developed.

The DB will accept data requests (queries) allowing the transfer of the appropriate data from the archive's storage volumes depending on user privileges. Archive users will browse the archive to access and retrieve CTA data of interest, they will query the archive database in order to select data based on specific selection criteria (source location, observation time interval, observation conditions criteria, etc). The high level products archive will provide Archive users with scientific products (images, spectra, light curves, catalogs, etc.) produced. Furthermore, these products will also be produced, archived and then distributed in a VO compliant format and accessed via dedicated VO data server(s).

Three levels of DB are identified within the CTAO context:

The Proposal Handling DB will allow the guest observer to retrieve the basic information on proposals as well as the status of the submitted ones. By means of this DB, the guest observer will be informed about the proposal ranking, the observations planning and the schedule with respect to the array's requested configuration and observability constrains.

The Archives DB will provide access to the data (EVT, CAL, MC) arranged within the hierarchy described in the previous section. It will allow CTA users to browse archive branches (including the MC branch) according to how the observations were performed, stored, analyzed and, eventually, linked together.

The Technical, engineering & monitoring DB will permit the retrieval of the TECH data describing the status of the CTA array subsystems including observation configurations, alerts and auxiliary information.

Complementary systems may be developed to distribute the high level DL3 data. The list of candidate gamma events could be accessible by guest observers by means of dedicated interfaces provided by the data access servers. Specific queries would be used to browse the event database and generate customized event lists. For example, a user might request all events with an energy larger than 50 TeV, for all publicly available observations. The given output (list of events and response data) would be provided following the appropriate standard to be used by the CTA Science Tools. Such a further service will be implemented or not depending on the final high level (DL3) data size that strongly affects both the management and the accessibility performance of the database.

2.3. Requirements estimation and associated challenges

Global data volume needs In order to handle a reasonable amount of data, it has been decided that not all RAW data generated by CTA's cameras will be conserved. Based on Monte Carlo simulations it has been estimated that data quality can be maintained keeping the the complete waveform of the signal in only 3% of the pixels. Even after this data reduction, the production of uncompressed RAW data is estimated to nearly reach 40 PB/yr. A reduction of a factor of 10 should be achieved on-site before data transfer to reach 4 PB/yr of RAW data. Such a compression ratio is not easily achieved and if not applied could cause many issues for the data archive. Some advancements on compression have been realised under the ASTERICS project and should be proposed as potential solution. The total amount of data generated per year is estimated at 11 PB, making an estimated total of 162 PB in 2031.

Computation needs Estimation of the CPU needs have been done considering the total wall time (sum of CPU time, I/O time and communication channel delay) and data requirements: one day of raw data acquisition must be processed in less than one day and the MC simulations and Instrument Response Function production to analyse one year of data must be processed in less than one month. The average core needs increase linearly from 1000 in 2017 to 9000 in 2031 (2013 CPU performance). Once a year, due to improvements in the reconstruction analysis software, all archived data must be reprocessed. The requirements impose that one year of data is reprocessed in one month. During this reprocessing period, the CPU core needs increase drastically ranging from 5000 to 9000 (2014 CPU performance) during the operation phase. This could represent a bottleneck for CTA computing model and efforts are being done to reduce the computing time of the analysis. Some advancements have been realised under the ASTERICS project and present potential solutions.

Network needs Data management requirements impose a maximum of 10 days to transfer daily raw data off-site. With a network bandwidth of 1 Gbps, a bandwidth efficiency of 80% and a data reduction factor of 10, it has been estimated that the requirement might be

exceeded after 7 sequential days of full observations for CTA South. Here again, this heavily depends on the data reduction factor achieved.

Distributed computing model The CTA computing model envisages four data-centers (DC), a number resulting from the trade-off between economy of scale and sustainability (compared to a centralized computing model). In this scenario, all computing tasks are distributed among the centers, in order to process data close to where they are physically stored. The main consequence is the increased complexity of the Archive Management System and File catalog associated, that will have the responsibility to manage this distribution following a strategy to be studied to optimize the data reduction and simulation pipelines. A coherent mapping of the dataset definition, distribution and storage, taking into account the frequency of access by users, has to be conceived for permanent storage.

Therefore, a performant distributed database system should be developed to handle efficiently the distribution of data between on-site analysis and off-site data centers. Moreover, the database should keep track of and synchronize all changes applied to raw data during the analysis chain (DL0 to DL5) and conserve information about the data provenance.

Currently, no official estimation on the size or query rates of such a database are available. Here again, insights from ASTERICS and other ESFRI projects might be extremely valuable.

3. SKA

3.1. Introduction

The Square Kilometre Array (SKA) will be the world's largest observatory. It will consist of two radio interferometers, located in South Africa and Australia with maximum baselines of order 120km in South Africa and of order 65km in Australia. These two instruments have complementary frequency coverage; in Australia the SKA-LOW instrument covers the frequency range 50-350 MHz, using log-periodic dipole antennas arranged into 512 "stations" and in South Africa the SKA-MID instrument will comprise 197 antennas (Incorporating the MeerKAT array) and operate in the 500MHz-15GHz frequency range.

SKA is divided into two phases, SKA1 and SKA2. SKA1 is due to begin construction in 2018 and to begin early science in 2020. Detailed design of SKA2 will begin in 2018 and thus SKA2 is not covered here [9].

3.2. Data Model

SKA presents a very challenging data environment due to its sheer scale. Due to the high data rates before processing, a shared processing facility for the two sites is not feasible. Furthermore, this means that post-processing must be done on each site – which is the remit of the SKA Science Data Processor (SDP) consortium [10].

Due to most of the processing being digitial, the SKA will be a highly flexible instrument, in which many different types of observations can be done. Science cases have been prioritized, and the main final data products are divied into two categories[11]: image cube type producs and UV-grid type products.

1. Imaging data for Continuum, as Taylor term images (images of the sky intensity and its derivative w.r.t. frequency). For Slow Transients detection have been specified - maps are made, searched and discarded)
2. Residual image (i.e. residuals after clean applied) in continuum Taylor terms.
3. Clean component image in Taylor terms (or a table, which could be smaller).
4. Clean component image for spectral line
5. Spectral line cube after continuum subtracted
6. Residual spectral line image (i.e. residuals after clean applied)
7. Representative Point Spread Function for observations

In addition, there are UV grid type products:

1. Calibrated visibilities, gridded onto grids at spatial and frequency resolution required by the observation. One grid per facet (so this grid is the FFT of the dirty map of each facet).

2. Accumulated Weights at each uv cell in each grid (without additional weighting (e.g. uniform) applied).

During data reduction, there are various data levels, in which the general idea is to discard as much information as possbile as soon as possible.

3.3. Archive size estimates

Archive size depends on the experiments performed, and with the one exception in the case of SKA-LOW (see below) there are no currently planned experiments that will drive data rates themselves. These archive sizes are the best current estimates of the SDP consortium [10].

Using the maximum baseline, a 50,000x50,000 pixels image could be produced with 216 frequency channels. At 8 bytes per pixel per channel, this would have a size of 1.3PB. Storing such data daily will lead to exabyte-scale archive requirements. For comparison, THINGS [12] - VLA images of nearby galaxies, released in 2008 - produced images 1024x1024 pixels in size that had 100 frequency channels and 4 bytes per pixel per channel. At 400MB these images could be easily stored and manipulated in desktop computers. If such full images are produced, they will be stored in the archive. However, not all observations will require images of this size. For the storage requirements, a weighted average has been taken over the different use cases [3].

For the LOW component of SKA, the size of the archive after 5 years is currently estimated to be 400PB, the majority of which will be from the EoR (Epoch of Reionization) observations - a custom experiment planned with LOW. This corresponds to a data rate out of the archive and to the external world of 22Gbit/s. Without EoR data, the 5 year archive is estimated to be 55PB with a data rate of 3Gbit/s.[11] For the MID component, the 5 years archive size is set to be 170PB, corresponding to a rate of 9Gbit/s.

3.4. Computing requirements

The computing requirements for SKA again depend on the actual science case. In the note SDP System Sizing [SKA-TEL-SDP-00000038], a weighted average is taken over the various use cases (HPSOs, High Priority Science Objectives), as the compute intensity of different use cases varies to a large extent. The total compute power required is measured in FLOPS: floating point operations per second. This is again converted to a requirement in terms of MegaWatt, since the maximum total power consumption of the SKA is fixed.

In this report, we will restrict ourselves to the Science Data Processor (SDP) part of the SKA. Before the signal reaches this stage, it has gone through a CSP (Central Signal Processor), which correlates or beamforms the data. The CSP will likely use FPGAs. We leave this part out of scope of this document because it uses different hardware and probably will not have much in common with the other projects.

4. LOFAR

4.1. Introduction

LOFAR is a telescope for observing the sky between 10MHz and 200MHz, with stations across Europe. It is a precursor to SKA, and particularly resembles SKA1 LOW, the low frequency part of SKA1 that will be built starting from 2018. The raw data rate of LOFAR is about 5 TB/s, whereas that of SKA1 LOW will be about 150 TB/s. LOFAR has many use cases such as interferometric imaging, pulsar searching, transient detection and the detection of cosmic rays. For this report we will focus on the use case of interferometric imaging.

4.2. Data model

4.2.1. Data levels

The data volume straight behind the A/D converters is so large that it is not possible to get all data to a central point. However, it is important to combine all data together to do interferometric imaging or beamforming. To achieve this, data is reduced / combined in various stages:

1. Data of 4x4 dipoles in the high band antennas is combined through analog beamforming.
2. On the station level, data is combined through digital beamforming (on FPGA boards). The result is a number of beams in a number of subbands. The number of beams and the number of subbands is limited by the data rate.
3. At the central correlator, data is correlated and integrated. After this, the data is temporarily stored on a cluster. Optionally, a copy of the data can be archived into the Long Term Archive (LTA) at this stage.
4. At high time and frequency resolution, radio frequency interference (RFI) is flagged, and data is partly calibrated to remove the contribution of bright sources outside the field of view. At the end of this optional pipeline, data can (again) be ingested into the LTA.

We will consider the offline processing (stage 4) as the stage that has the most in common with the other instruments, and will focus on the challenges of that part.

The offline processing is performed on a dedicated compute cluster at the same site of the correlator. The correlator dumps its data straight onto the disks of the compute cluster. Indeed, this cluster can be seen as a semi-online system, since data only resides there for a short time. The compute cluster (“CEP4”) consists of ~50 CPU machines and 4 GPU machines. The GPU machines are intended to be used for imaging, while the CPU machines are used for flagging, calibration and averaging. The nodes are connected through an infiniband network, and share a single Lustre filesystem.

The software that performs calibration and imaging is mainly developed in-house, but also uses external packages. Files are stored in the domain specific ‘casacore’ framework, which is common to all of radio astronomy.

To handle changing versions and dependencies, pipelines are run in a Docker container. This has the advantage that several different environments can run side by side, and that the software environment is versioned.

Most processing is embarrassingly parallel, since the processing can be done on different (groups of) subbands. Typically the number of subbands of an observation is about 400. Managing the jobs is done by a workflow management system developed by Astron, but there are plans to replace this by a general purpose workflow management system.

5. EUCLID

5.1. Introduction

EUCLID is an ESA mission aiming at understanding the nature of the dark energy and dark matter by accurately measuring the accelerated expansion of the Universe. It is composed of one space telescope located at L2 (the second Sun-Earth Lagrange point) making a wide extra-galactic sky survey in optical and near-infrared and supported by complementary observations from ground telescopes. The launch is planned in late 2019.

5.2. Data model

EUCLID carries two instruments, an imager in the visible domain (VIS) and an imager-spectrometer (NISP) covering the near-infrared that will generate 850 Gbit/day of compressed science data. In order to have full control over the sources systematic errors, no VIS image processing will be done on board and all CDD data are transferred to the ground. NISP does the detector and voltage ramp processing on board providing one photometric or spectroscopic image per exposure. The science data are transferred to ground by using K-band (26 GHz) transponder.

5.2.1. Data levels

The Euclid data processing system is organized in sequential processing steps of increasing sophistication. With each step is associated a data processing level or “data level”. Data levels consist of all data produced by the corresponding data processing step including intermediate data. The data levels are as follow:

Level 1 data: raw VIS and NISP images.

Level 2 data: calibrated and co-added images from VIS and NISP, validated for cosmology analysis; PSF model and optical distortion maps; co-added spectra.

Level 3 data: catalogs (including redshift, ellipticity, shear, etc.); dark matter mass distribution; shear and galaxy correlation functions and covariance errors; additional science catalogs; ground based information which was used in the derivation of the data products.

Transients: transient events data products, including the derived transient category (e.g. supernova, solar system object, etc.) and brightness, target position and possible finding chart.

Level Q data: products defined so that they are suitable for most purposes in Astronomy, except for the main cosmological goals of the mission.

Level E data: quality-controlled external data from existing missions and ground-based surveys which are used for calibrations and photometric redshift derivations

Level S data: pre-launch simulations and modeling impacting on calibrations and observing strategies

5.2.2. Data centers

The data are collected, archived and processed locally by distributed Science Data Centers (SDC) following the idea of "moving the code, not the data". Each SDC is both a processing and a storage node. There is separation of metadata from data: a centralised metadata repository will be available (at SOC), containing "pointers" to the actual pixel data distributed geographically across the SDCs. For integrity purposes, the metadata repository and the bulk data shall be mirrored.

There will be no SDCs dedicated to specific tasks. SDCs are considered as generic resource providers capable of providing the services requested by the SGS needed. As a consequence, any pipeline will be able to run on any SDC: therefore, each SDC runs the same code through virtualization techniques.

5.2.3. Data processing

Different operational units (OU) are in charge of processing the different levels of data as well as different kind of data. The SOC (Science Operation Center) deposits the Level 1 data in the Euclid archive system. Three OUs, responsible for the calibration of the data, develop pipelines to carry out the instrument specific processing which yields the Level 2 data. Other OUs are responsible for the processing of the ground based data (OU-EXT, level E), the merging of the data (OU-MER), and simulations (OU-SIM, level S). The OUs SPE, SHE, and PHZ process the level 2 data to obtain spectroscopy, galaxy shear and photo-z catalogs. These catalogs are further processed by the OU-LE3 to obtain the level 3 catalog.

Processed data are associated with quality control information that ensures traceability of input data sets as well as the processing steps applied.

All intermediate and final data-set and associated quality control and processing information are stored into the Euclid Mission Archives (EMA). The EMA constitutes the "working" repository of the mission and is used for disseminating data within the Euclid collaboration. The Euclid Legacy Archive (ELA) will provide access to the final validated products to the general scientific community.

5.3. Resource estimation and associated challenges

The logical internal architecture of the Euclid SGS is based on four main pillars: (1) an *Euclid Archive System (EAS)*, composed of a single Metadata Repository which inventories, indexes and localizes the distributed data; and a Distributed Storage of the data over the SDCs

(ensuring the best compromise between data availability and data transfers), with some redundancy; (2) A set of *Services* which allow a low coupling between SGS components, e.g.: metadata query and access, data localization and transfer, data processing monitoring and control; (3) An *Infrastructure Abstraction Layer (IAL)* allowing the data processing software to run on any SDC independently of the underlying IT infrastructure, and simplifying the development of the processing software itself (e.g. takes care of interfaces); (4) A *Monitoring & Control and Orchestration Layer* responsible for distributing data and processing among the SDCs.

Euclid mission is divided into two phases: pre-launch and post-launch. Of course the amount of data will significantly increase post-launch but significant numeric resources are also required pre-launch for the simulation providing the necessary data to test and develop the analysis pipeline. The resources requirements post-launch should increase horizontally with the data acquisition. However, the resource requirements pre-launch evolve with time and are planned to peak in 2019 (next peak values are given for 2019).

Processing budget

Resources requirements pre-launch are divided between simulation and computation (for validation of the analysis purposes). The peak storage needs is evaluated at about 16PB (respectively 1PB for simulations and 15PB for computation) and peak computing needs at 2200 core*year.

Post launch resources grow linearly with time. Storage needs are evaluated to go from 10PB in 2021 to 100PB in 2027. Computing needs are evaluated to go from 3000 core*year in 2021 to 21000 core*year in 2027.

Data Transfer budget

The main bottleneck for data transfers concerns the transfer of the data products to SOC for long term archive. The data release is planned to be 5.3 PB for DR1, 14.9 PB for DR2 and 31,9 PB for DR3. With a rate of 2.5 Gb/s, DR1 only will require 200 days to be transferred to the long term archive. As this is not realistic, two alternatives are being considered: a reduction factor of at least 10, achieved by removing raw exposures or using the SDC's storage infrastructure as long term archive.

Database

Euclid catalog should contain more than 10^{10} objects (with ~1500 parameters characterizing each object), imposing huge constraints on the database.

The spectra 1D and 2D of sources should be implemented in files (probably FITS format) and some metadata will describe these spectra in the Euclid metadata database. The volume of these metadata goes from 1KB to 3.8MB per source depending on the instrument (on board or on ground). With a number of sources per observation ranging from 1500 to 240000, this

sums up to a total of 56GB per observations to be ingested into the Euclid Archive metadata base. The number of observations will grow linearly (from 21000 in 2021 to 126000 in 2027) in post-launch phase. Multiplying those inputs gives the average catalog throughput per hour. The result goes from 135 GB/hour in 2021 to 808 GB/hour in 2027, showing the tremendous demand on I/O of the metadata base (only for metadata ingestion). Therefore, the catalogs should be stored as files in the SDC, as closely as possible to the processing infrastructure.

6. KM3NeT

6.1. Introduction

KM3NeT, located in the abysses of the Mediterranean Sea, is a distributed research infrastructure that will host a km3-scale neutrino telescope (ARCA), offshore from Capo Passero in Italy, for high-energy neutrino astronomy, and a megaton scale detector (ORCA), offshore from Toulon in France, for the determination of the neutrino mass hierarchy.

For both cases, the detector arrays comprise a three dimensional grid of photomultipliers tubes designed to detect the Cherenkov light induced by charged leptons produced by neutrino interactions in and around the instrumented volume. KM3NeT has developed a cost effective Optical Module based on many small 3" photomultiplier tubes. Depending on the neutrino energy range of interest, the Optical Modules are configured in a dense (ORCA) or sparse (ARCA) geometry. Recently, the first KM3NeT detection strings have been successfully deployed and are providing high quality data. The construction of the infrastructure will be completed by 2020.

6.2. Data Policy [3]

The KM3NeT Collaboration has developed a data policy based on the research, educational and outreach goals of the facility. The first exploitation of the data is granted to the collaboration members as they build, maintain and operate the facility and to priority users. Accordingly, each collaboration member has full access rights to all data, software and know-how. Access for non-members is restricted, as long as methods and results have not yet been published. The prompt dissemination of scientific results, new methods and implementations is a central goal of the project, as is education. High-level data (event information enriched with quality information) will be published after an embargo time of two years under an open access policy on a web-based service. Exceptional access rights that correspond to these goals can be granted.

The Collaboration has developed measures to ensure the reproducibility and usability of all scientific results over the full lifetime of the project and in addition 10 years after shutdown. Low-level data (as recorded by the experiment) and high-level data will be stored in parallel at central places. A central software repository, central software builds and operation system images are provided and continuously maintained until the end of the experiment.

The storage and computing needs of the KM3NeT project are highly advanced. The Collaboration has developed a data management plan and a corresponding computing model to answer those needs. The latter is based on the LHC computing models utilising a hierarchical data processing system with different layers (tiers).

Tier	Computing Facility	Processing steps	Access
Tier-0	at detector site	triggering, online-calibration, quasi-online reconstruction	direct access, direct processing
Tier-1	computing centres	calibration and reconstruction, simulation	direct access, batch processing and/or grid access
Tier-2	local computing clusters	simulation and analysis	varying

Data are stored on two main storage centres (CCIN2P3-Lyon, CNRS and CNAF, INFN); those large data centres are fully interfaced with the major European e-Infrastructures, including GRID-facilities (ReCaS, HellasGRID provide resources to KM3NeT). The main node for processing of the neutrino telescope data is the computer centre in Lyon (CCIN2P3-Lyon). A corresponding long-term and sustainable commitment has already been made by CNRS, which is consistent with the needs for long-term preservation of the data. A specialised service group within the Collaboration will process the data from low-level to high-level and will provide data-related services (including documentation and support on data handling) to the Collaboration and partners. WAN (GRID) access tools (e.g. xrootd, iRODS, and gridFTP) provide the access to high-level data for the Collaboration. The analysis of these data will be pursued at the local e-Infrastructures of the involved institutes (both local and national). The chosen data formats allow for the use of common data analysis tools (e.g. the ROOT data analysis framework) and for integration into e-Infrastructure common services.

The central services are mainly funded through CNRS and INFN that have pledged resources of their main computing centres to the project. Additional storage space and its management are provided by the partner institutes.

In addition to the major storage, networking and computing resources provided by the partner institutions and their computing centres, grid resources have been pledged and will be used by KM3NeT (ReCaS, HellasGRID). These will provide significant resources to be used for specialised tasks (as e.g. for special simulation needs). The major resources, however, will be provided by the partners. External services are employed to integrate the KM3NeT e-Infrastructure into the European context of the GRID – in the fields of data management, security and access; services will be implemented in collaboration with EGI.

One of the aims of the KM3NeT data management plan is to play an active role in the development and utilisation of e-Infrastructure commons. KM3NeT will therefore contribute to the development of standards and services in the e-Infrastructures both in the specific research field and in general. In the framework of the Global Neutrino Network (GNN), KM3NeT will cooperate with the ANTARES, IceCube and GVD collaborations to contribute to the open science concept by providing access to high-level data and data analysis tools, not only in common data analyses but also for use by citizen scientists.

In the framework of the ASTERICS project, KM3NeT will develop an interface to the Virtual Observatory including training tools and training programmes to enhance the scientific impact of the neutrino telescope and encourage the use of its data by a wide scientific community including interested citizen scientists. Data derived from the operation of the experiment (acoustics, environmental monitoring) will be of interest also outside of the field. Designated documentation and courses for external users will therefore be put in place to facilitate the use of the repositories and tools developed and used by the KM3NeT Collaboration.

6.3. Requirements estimation

For phase 2.0 of KM3NET, it is estimated that 2500 TB of storage will be required. The required computing time is estimated at 109 HS06.h. The total computing resources are estimated at 125.000 HS06.

7. LSST

7.1. Introduction

The Large Synoptic Survey Telescope (LSST) [1], currently under construction in Chile, is designed to conduct a ten-year survey of the dynamic universe. This large-aperture, wide-field, ground-based telescope will map the entire southern sky in just a few nights in six optical bands from 320 to 1050 nm with its 3.2-gigapixel camera. LSST will take about 2000 exposures per observing night, for a total raw data volume of 15 TB per 24 hours period. Detected and measured objects will be stored in a database catalog that, in its final year, is estimated to include 26 billion stars and galaxies in dozens of trillion detections forming a multiple petabytes dataset.

7.2. Data Model

The LSST data products are organized into two groups, distinguished by the frequency with which they are generated. Divided into images, catalogs, and alerts, the level 1 products are generated by pipelines processing the stream of data from the camera system during normal observing periods, and are therefore being continuously generated and updated every observing night. Level 2 products, including calibration images, co-added images, and the resulting catalogs, are generated on a yearly basis. Level 3 products are under the responsibility of the science collaborations that will exploit the LSST dataset.

7.2.1. Data levels

The observatory, telescope, camera and data management systems are specifically designed to conduct the survey and will deliver three levels of data products and services.

Level 1: Nightly data products. They will include images, difference images, catalogs of sources and objects detected in difference images, and catalogs of Solar System objects with their associated orbital elements. Their primary purpose is to enable rapid follow-up of time-domain events. The catalogs will be entered into the Level 1 database and made available in near real time. Notifications about new sources will be issued using community-accepted standards within 60 seconds of observation. Several millions of alerts are expected to be generated each night.

Level 2: Annual data products. They will include well calibrated single-epoch postage-stamp images, deep co-additions, and catalogs of objects, sources, and forced sources, enabling static sky and precision time-domain science. It will also include fully reprocessed Level 1 data products. In contrast to the Level 1 database, which is updated in real-time, the Level 2 databases are static and will not change after release.

Level 3: User-created data product services. They will enable science cases that greatly benefit from co-location of user processing and/or data within the LSST Archive Center. Recognizing the diversity of astronomical community needs, and the need for specialized processing not part of the automatically generated Level 1 and 2 products, LSST plans to devote 10% of its data management system capabilities to enabling the creation, use, and federation of Level 3 data products.

7.2.2. Archives

Over 10 years of operations, LSST will produce eleven data releases: two for the first year of survey operations, and one every subsequent year. Each data release will include reprocessing of all data from the start of the survey, up to the cutoff date for that release. The content of data releases is expected to range from a few PB for DR1 to 70 PB for DR11. These data releases will include the raw postage-stamp images, retained co-adds, and catalogs. Given that scale, it is not feasible to keep all data releases loaded and accessible at all times. Instead, only the contents of the most recent data release, and the penultimate one will be kept on fast storage and with catalogs loaded into the database. Older releases as well as raw exposures will be archived to mass storage (tapes). The users will not be able to perform database queries against archived releases. They will be made available as bulk downloads in some common format. All raw data used to generate any public data product (raw exposures calibration frames, telemetry, configuration metadata, etc.) will be kept and made available for download.

LSST facilities include a mountain summit/base facility, a main central Archive Center at NCSA[1] (US) associated to a remote Satellite Archive Center at CC-IN2P3[2] (France), multiple Data Access Centers, and a System Operations Center. The data will be transported over high-speed optical fiber links from the mountain summit/base facility in South America to the Archive Center in the U.S and in France. Data will also flow from the mountain summit/base facility and the Archive Centers to the Data Access Centers over existing fiber optic links.

The Archive Centers are super-computing class data centers with high reliability and availability. This is where the data will undergo complete processing and re-processing and where they will be permanently stored. The Archive Centers also constitute the main repositories feeding the distribution of LSST data to the science community. Following an agreement between CNRS/IN2P3 and LSST the CC-IN2P3 satellite archive center will be in charge of processing 50% of the level-2 data. A complete copy of the data (raw + catalogs) will be available at CC-IN2P3, along with the corresponding level-1/3 data and catalogs. The LSST infrastructure at CC-IN2P3 will be fully integrated to the LSST Data Release Production system. While a full compatibility is required between the centers, they will not necessarily share the same hardware and middleware.

A network of data access centers is envisioned for broad user access.

7.2.3. Database management system

The guiding principle of the LSST project is that the vast majority of science cases covered by LSST should be enabled by catalog queries without having to go back to images. Indeed, these catalogs will contain the physical properties of the observed astronomical objects that will be automatically measured by widely accepted and published algorithms, all included in the LSST data management stack [2]. Therefore, access to images by the individual scientists should be rarely needed. The catalog of properties of celestial objects is at the core of the data management system built for LSST to reach its scientific goals.

To satisfy the need to efficiently store, query, and analyze these catalogs, that will ultimately contain trillions of rows and petabytes of data, the LSST database team, based at SLAC National Accelerator Laboratory (Stanford) and with contributions from CNRS/IN2P3, are building a prototype system for user query access, called Query Service (Qserv)[3][4], an open source distributed shared-nothing SQL database system. The system relies on several production-quality components, including MySQL and XRootD[3] / Scalla[5]. The key requirements driving the LSST database architecture include incremental scaling, near real-time response time for ad-hoc simple user queries, fast turnaround for full-sky scans/correlations, reliability, and low cost, all at the multi-petabyte scale. Qserv is currently under active development and is also tested on real data sets. Thanks to a partnership with Dell, a first Qserv test bench has been deployed with CC-IN2P3 for large scale tests (50 nodes - 400 cores - 800 GB memory - 500 TB disk storage), and is currently replicated at NCSA. The current tests include 35TB of data, while the next step will contains 120TB from the Wide-field Infrared Survey Explorer (WISE) data set that will be used to populate the database.

7.3. Requirements estimation

Requirement estimations are presented and summarized in the last section of this document in one table and two figures, along with the other projects.

8. Comparison between projects

The table below shows some key facts for each instrument; the numbers are only indicative.

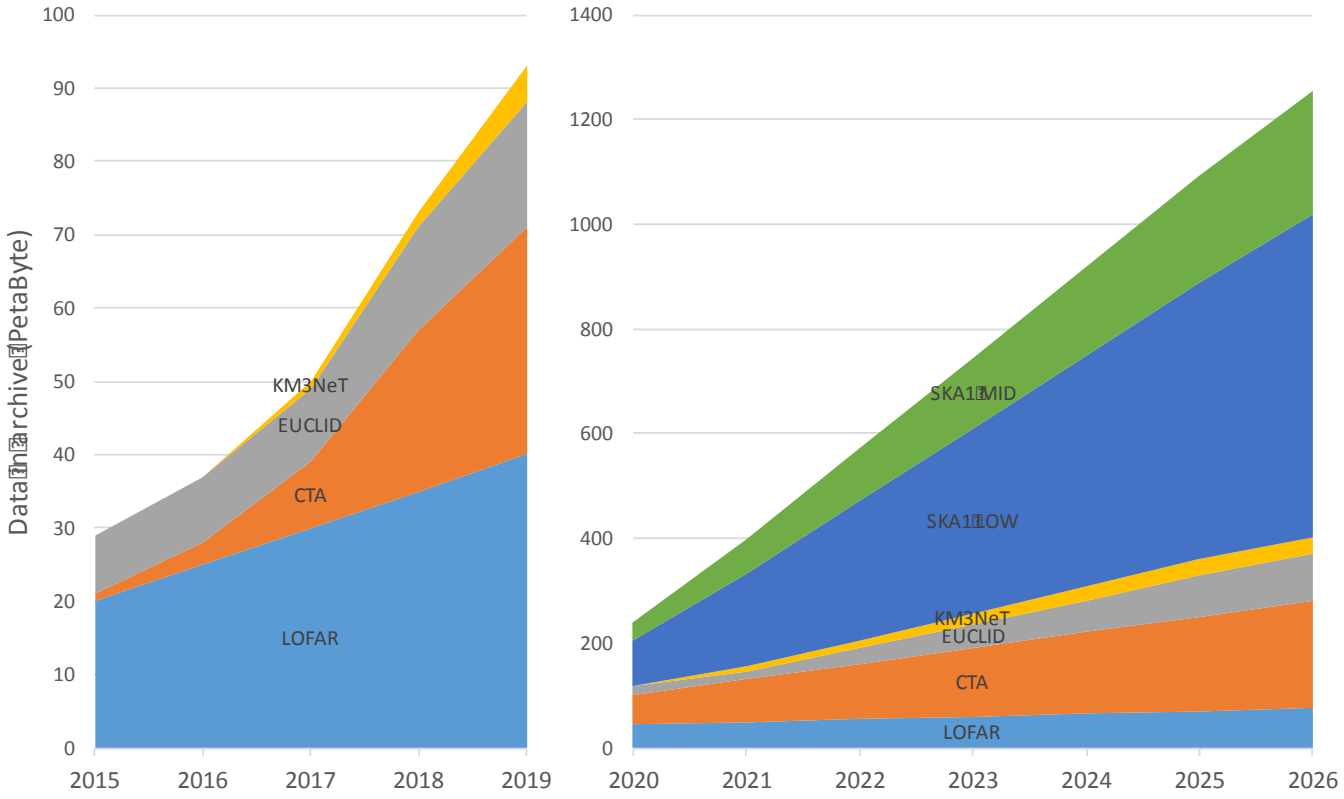
	LSST	LOFAR	SKA1 LOW	SKA1 MID	CTA	KM3NeT Ph. 2.0	Euclid
Type of data	Images	Images	Images	Images	Events	Events	Images
Raw data rate	1 GB/s	5 TB/s	150 TB/s	2 TB/s	7.8 GB/s	40 GB/s	0.02GB/s
Data rate to archive	5.5PB/year	5 PB/year	77 PB/year (TBC)	34 PB/year (TBC)	25 PB/year	1-3 PB/year	7 PB/yr
Observation Duty cycle	50%	70%			10%	>98%	>98%
Number of stations	1	50	500	200	118	3	1
Total number of detection elements	1 camera of 3.2GPx	100.000 antennas	130.000 antennas	~200 dishes	1 camera of <10 KPx per station	192.510	2 imagers
Data products	• Raw data • Calibrated images • Catalogs	• Raw data • Calibrated data • Images • Catalog	• Raw data • Calibrated data • Images • Catalog	• Images • Catalog	• Raw data • Calibration data • Technical data • Monte Carlo data • High level Multi-frequency archive	• Raw data • Calibration data • Simulation data • High-level science data • Observer data archive • Auxilliary and earth/science data	• Raw data • Calibrated images • Simulation data • Catalogs • Transients • Other science data
Files to archive	270.000/night	50 files / day	~100-100/day TBC	~100-100/day TBC	TBD	~20 files/day	TBD

Comparison between projects

	LSST	LOFAR	SKA1 LOW	SKA1 MID	CTA	KM3NeT Ph. 2.0	Euclid
Archive	Distributed over 3 sites	Decentralized (grid)	On site archive, with mirroring at regional centres	On site archive, with mirroring at regional centres	Decentralized (grid)	Decentralized (grid, central computing centres CNAF and CC-IN2P3)	Distributed on 9 Data Centers
Archive structure	Files + DB for metadata	Files, oracle DB (Astro-WISE) for metadata	TBD	TBD	Database for files localisation and metadata	Files and oracle DB for metadata	Distributed + centrale database
Archive media	Archive: Files on tapes, staging on disk ; Catalog: DB (Qserv)	Files on tape, staging on disk	TBD	TBD	Files on tape	Files on tape, staging on disk	Files on tapes, database on disk
Archive interface	www (FireFly)	www, srm, gridftp	TBD	TBD	gridftp	www, srm, gridftp, iRODS	TBD
Computing needs							

9. Evolution of data storage

The following chart represents the evolution of the total storage needs for the different ESFRI projects in the coming years. Data sizes are in PetaBytes (1 PetaByte = 1.000 TeraByte). Note that a different scale has to be used starting 2020 when data volumes are foreseen to be very large.



10. References

[1] National Center for Supercomputing Application (Illinois)

[2] Computing Center of the National Institute for Nuclear Physics and Particle Physics (CNRS - Lyon - France)

[3] http://xrootd.org/

[4] “LDM-135: Database Design” - Jacek Becla et al. - http://ldm-135.readthedocs.io/en/master/

[5] “Scalla: Structured Cluster Architecture for Low Latency Access” - Andrew Hanushevsky and Daniel L. Wang - http://xrootd.org/papers/Scalla_IPDPS12.pdf

[6] CTA Data Management Technical Design Report

[7] T. Berghöfer, et al. (APPEC), Towards a Model for Computing in European Astroparticle Physics, http://www.appec.org/towards-a-white-paper.html

[8] S. Adrián-Martínez, et al. (The KM3NeT Collaboration), Letter of intent for KM3NeT 2.0, Journal of Physics G: Nuclear and Particle Physics, 43 (8), 084001, 2016, arXiv:1601.07459 [astro-ph.IM]

[9] http://www.skatelescope.org/wp-content/uploads/2013/08/SKA-Timeline-Gantt-Chart.png

[10] http://www.ska-sdp.org

[11] SDP (SKA Data Processor) document number SKA-TEL-SDP-0000038

[12] The HI Nearby Galaxy Survey (THINGS) http://www.mpia.de/THINGS/Overview.html