Data Management challenges in Astronomy and Astroparticle Physics

Giovanni Lamanna
LAPP-IN2P3-CNRS

4th LSDMA SYMPOSIUM – The challenge of Big Data in Science
KIT, 1 October 2015

OUTLINE

-> Astronomy is experiencing a deluge of data

-> The new ASTERICS-H2020 project

ASTRONOMY and ASTROPARTICLE

During the last decade:

- intensive construction of large astroparticle physics experiments and detectors.
- new perspectives in Astronomy and new infrastructures in preparation.

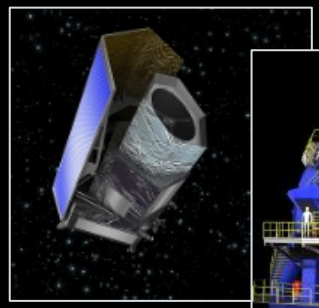
Towards the end of the decade:

- the projects passed from the noise hunting regime to the generation of large sets of data. Data production needs large computing resources, intensive simulation and large storage space.
- multi-messengers data need formats, software and services for wide accessibility and effective mining.

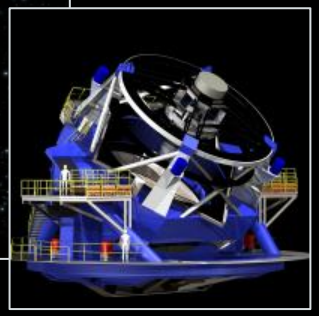
Radio Infrared **Visible light** X-rays **Gamma rays**

LOFAR

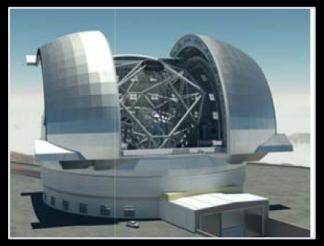
SKA



EUCLID

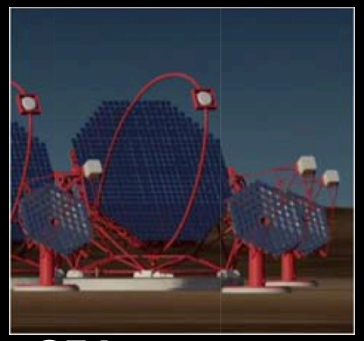


LSST



E-ELT

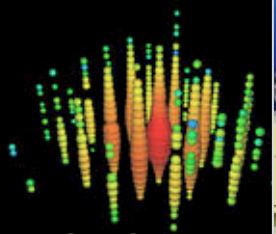
HESS



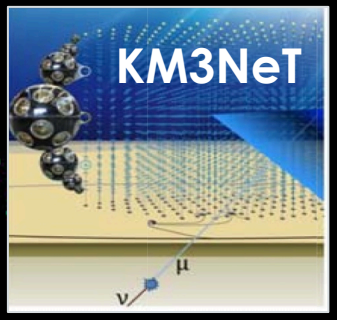
CTA

Gravitational Waves Cosmic-rays **Neutrinos**

LIGO & VIRGO



ICECUBE

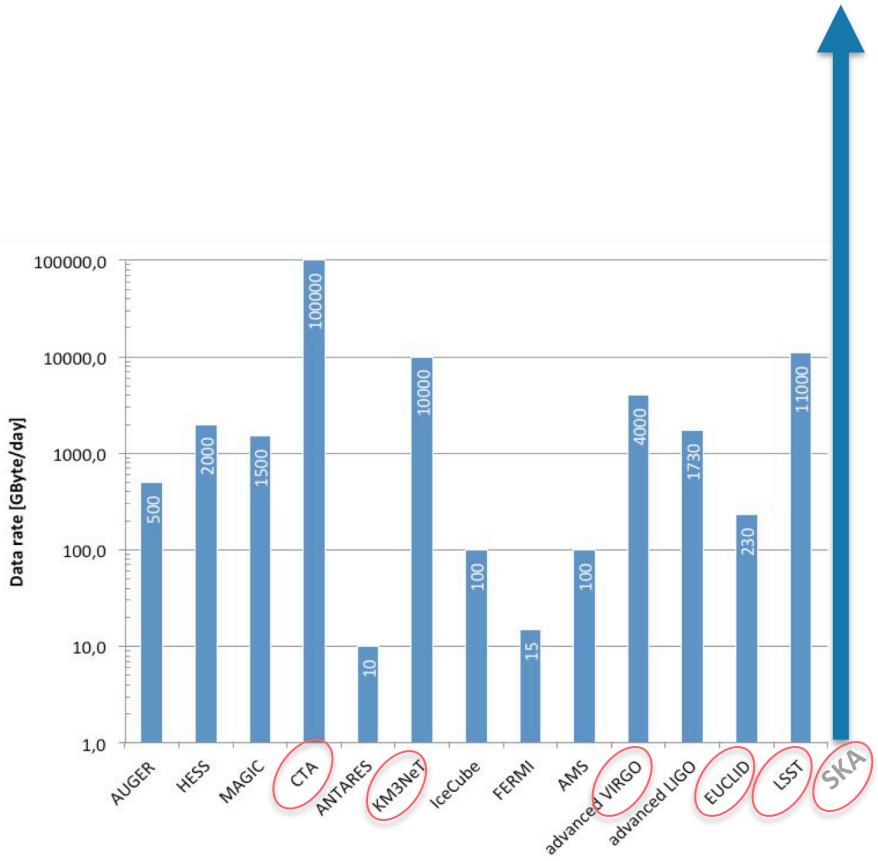


DATA RATES

Delivered data rates range from 10–100 GBytes/day to a few TBytes/day and up to 100 TBytes/day for future high-energy gamma-ray observatory (CTA), 15 TBytes per day for large telescopes (LSST*).

Altogether a data rate from Astroparticle projects of PBytes/year (~ LHC output).

The radio-astronomy project SKA** has rates extremely high: 160 TB per second raw data (in the first phase) …



Data rates by currently running and planned research infrastructures

* *In ten-year survey lifetime, LSST will map tens of billions of stars and galaxies*

** *The dishes of the SKA will produce 10 times the global internet traffic.*

COMPUTING…

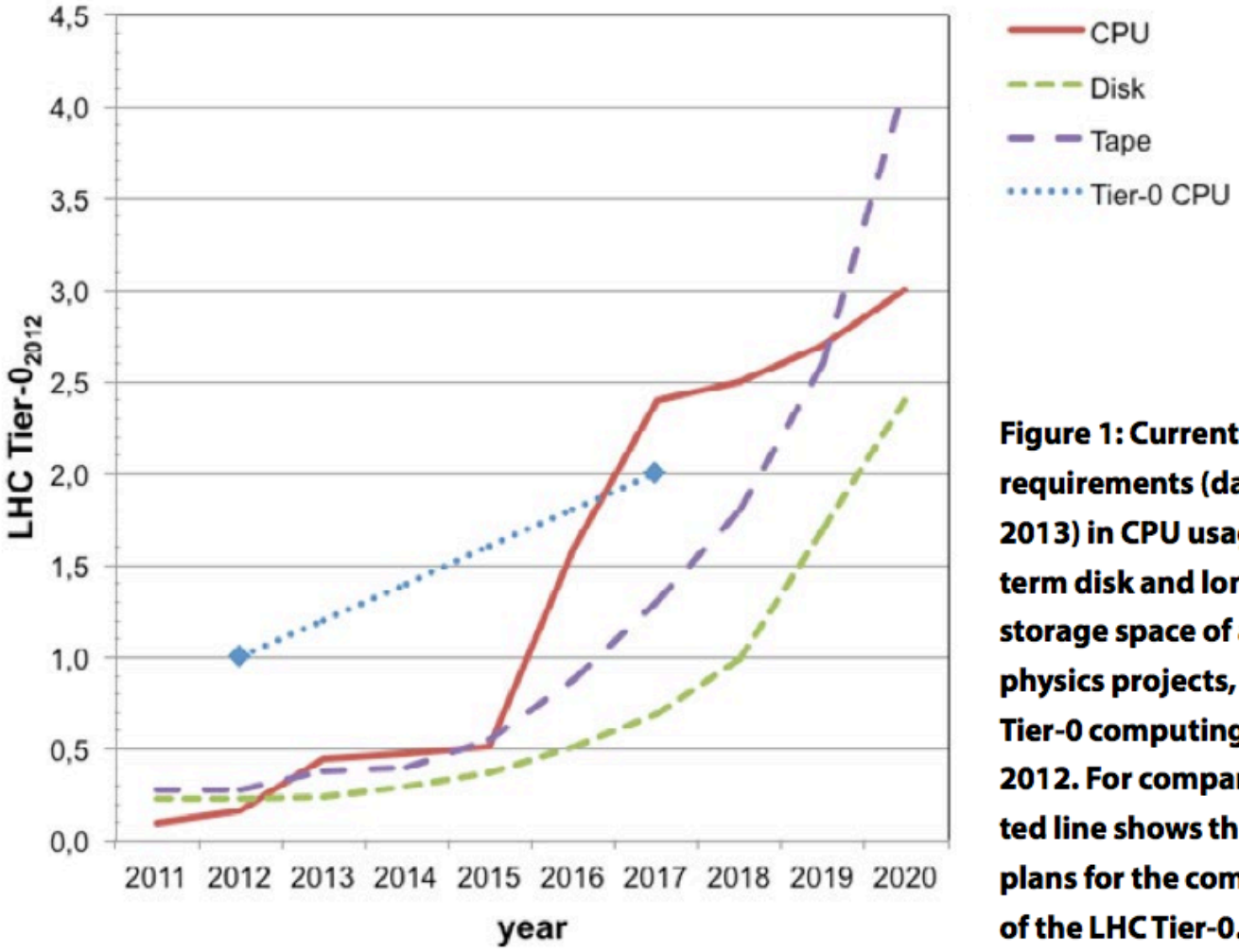


Figure 1: Current use and future requirements (data collected in 2013) in CPU usage and short term disk and long-term (tape) storage space of astroparticle physics projects, in units of LHC Tier-0 computing and storage in 2012. For comparison the dot-ted line shows the extension plans for the computing power of the LHC Tier-0.

Different probes/methods/specifications

Projects	Processing	Main requirements/challenges
EVENT-BASED *(γ-rays, CR, ν)*	Evt-builder, calib. and reconstruction; reduction, real-time science.	Raw big-data (storage & HTC centres). Data formats. Algorithms. On-site operation and reduction. Cooperative science tools. Observatory (A&A). Multi-λ.
IMAGE-BASED *(far-IR, VIS)*	Surveys/deep observation; combining photometer and spectrograph info.; Catalogue of objects.	Big-data products: data base challenges. Graphical processing, Algorithms. Images format. Catalogue preservation and query. HTC centres.
SIGNAL-BASED *(Radio, GW)*	Noise cleaning; mathematical processing (FT) converting signal in images.	Algorithms. New computing architectures. HPC and HTC combined. Fast soft reduction. Data mining and preservation.

Commons and cooperation:

ASTERICS

Astronomy ESFRI & Research Infrastructure Cluster
ASTERICS - 653477

- Astronomy ESFRI & Research Infrastructure Cluster
- Horizon 2020 Work Programme INFRADEV-4-2014/2015 Call – "Implementation and operation of cross-cutting services and solutions for clusters of ESFRI and other relevant research infrastructure initiatives"
- Focus of ASTERICS: SKA, CTA, KM3NeT, close links to E-ELT, EGO, EUCLID, LSST.
- Funded at 15 M€ for 4 years
- 22 partners in 6 countries, representing a major collaboration in Astronomy/Astrophysics/Astroparticle Physics
  ASTRON, CNRS, INAF, UCAM, JIVE, INTA, UEDIN, UHEI, OU, FAU, VU, CEA, UVA, UGR, FOM, IEEC, IFAE, UCM, INFN, STFC, DESY, SURFnet.

ASTERICS scope

- Bring together astronomy and astroparticle physicists:
- ESFRI facilities:
  - Radio (SKA)
  - γ-Ray (CTA)
  - ν (KM3NeT)
  - Optical (E-ELT)
- Aspiring ESFRI facilities, e.g. Einstein Telescope
- Complementary facilities, e.g. LOFAR, Euclid, LSST, VIRGO, LIGO, eVLBI, HESS, MAGIC, ANTARES, IceCube, etc.
- "Gathering critical mass… common solutions… cross-cutting activities… complementarity… interoperability… economy of scale"

ASTERICS overview

Astronomy ESFRI & Research Infrastructure Cluster
ASTERICS - 653477

WP1- AMST: ASTERICS **M**anagement **S**upport **T**eam

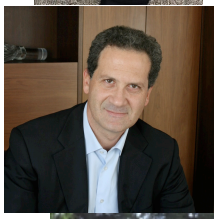
Coordinator: M. Garret

WP2- DECS: Dissemination, **E**ngagement and **C**itizen **S**cience

Lead: S. Serjeant

WP3- OBELICS: Observatory **E**-environments **Li**nked by common **C**hallenge**S**

Lead: G. Lamanna



WP4 - DADI: Data **A**ccess, **D**iscovery and **I**nteroperability (Virtual Observatory)

Lead: F. Genova

WP5 – CLEOPATRA: Connecting **L**ocations of **E**SFRI **O**bservatories and **P**artners in **A**stronomy for **T**iming and **R**eal-time **A**lerts.

Lead: A. Szomoru

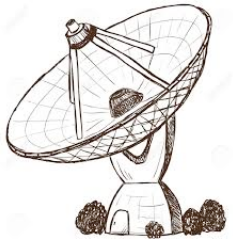
WP3 – OBELICS

OBservatory **E**-environments **LI**nked by common **C**hallenge**S**

- The ASTERICS core work package.
- Targeting common ESFRI-projects « Data Challenges ».
- Scopes:
  - Enable interoperability and software re-use.
  - Enable open standards and software libraries for multi-messenger data.
  - Develop common solutions, share prototypes, exchange experience.
- Expected impact:
  - Economies of scale and saving resources.
  - Contribute to the construction and operation of ESFRI projects.

WP3-OBELICS

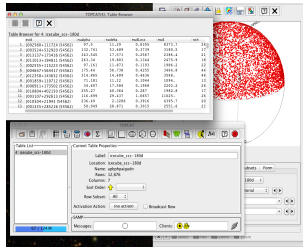
Working on commons along the "data flow":



1. Data generation

2. Data systems integration

3. Data analysis

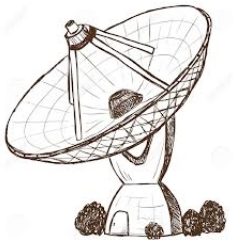


Twelve international partners cooperating around three main steps of data pipelines of major ESFRI projects in Astronomy.

OBELICS TASKS:

1. **D-GEX: Data GEneration and information eXtraction**
2. **D-INT: Data systems INTegration**
3. **D-ANA: Data ANAlysis/interpretation**

OBELICS commons

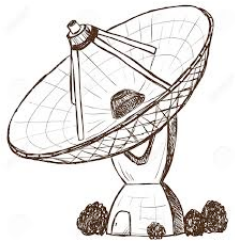


1. Data generation

D-GEX: Data GEneration and information eXtraction:

- ✓Surveying real-time or close-to-detector data streaming frameworks.
  (e.g. Hadoop, ACS and others; aiming at file and metadata management, fast algorithms integration, automatic remote acquisition, identification and ingestion..)
- ✓Standards on data model and data format.
  (e.g. Protocol buffer saving bandwidth; HDF5 simplifying big-data structure; evolution of scientific data FITS format; streaming protocols adopted in space projects...)
- ✓Prototype libraries handling secondary data streams.
  (environmental and engineering data, temporary local archive, device control software and observation scheduling)
- ✓Benchmarking low-power computer platforms. *(GPU + ARM, FPGA, Microservers, …)*

OBELICS commons



1. Data generation

2. Data systems integration

D-INT: Data systems INTegration

- ✓ Scaling-up existing databases and storage architectures beyond the Peta-scale level for complex queries.

  (Cooperative activities on "identification and archiving interesting data products")

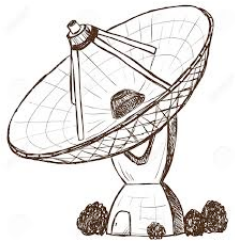
  i) Developing prototype benchmarks of large size DB: Cassandra, MongoDB, Qserv.

  ii) Testing and adopting data-management- system services for data-sets integration: FLUME, RUCIO, …

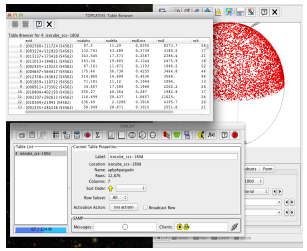
  iii) Multi-parameter instrument response function integration.

OBELICS commons

Astronomy ESFRI & Research Infrastructure Cluster
ASTERICS - 653477



1. Data generation → **2. Data systems integration** → **3. Data analysis**

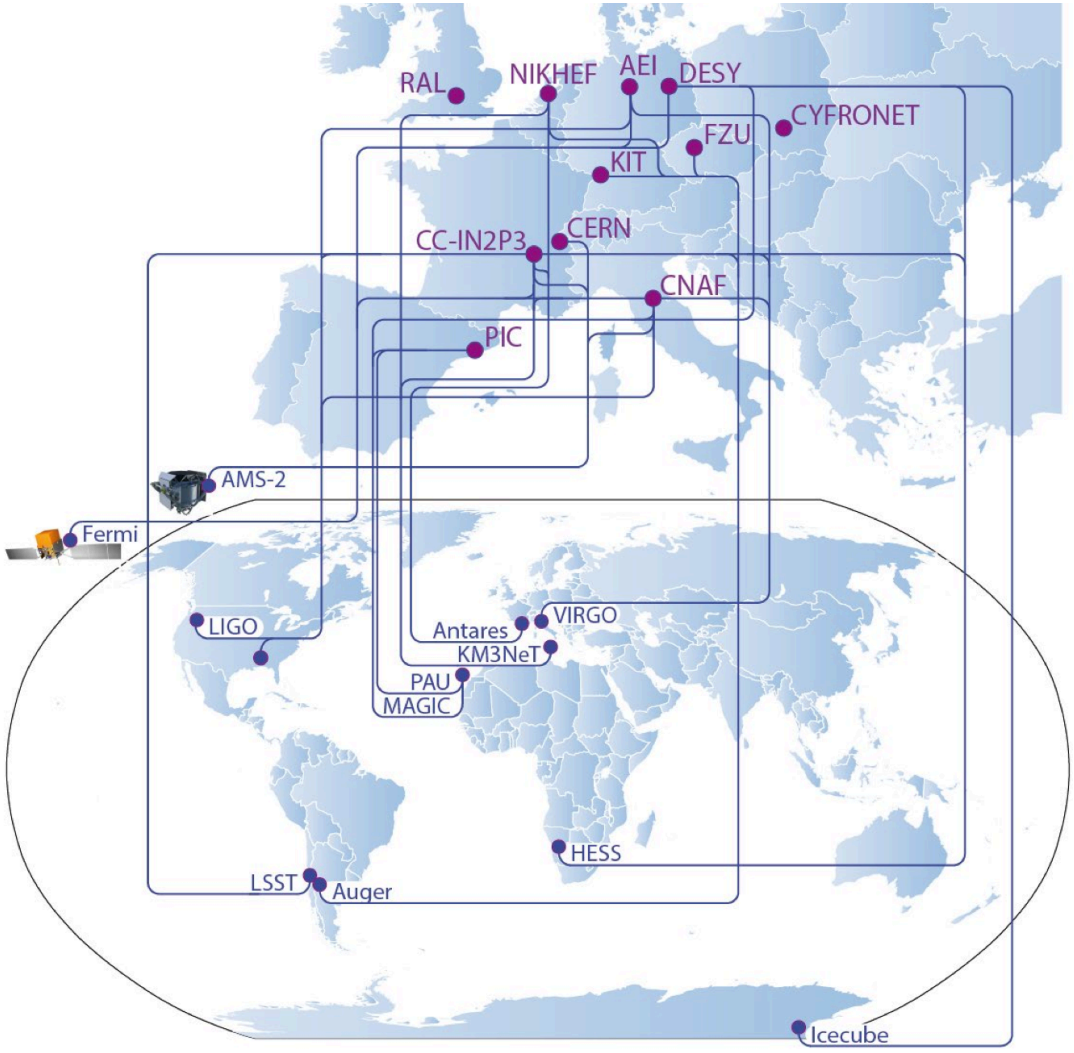


D-ANA: Data ANAlysis/interpretation

- ✓ Open source software for data analysis.
  - *a) Bayesian and likelihood analyses approaches for cross-matching between catalogues and transients detected via different instruments.*
  - *b) Simultaneous feature classification and extraction in multi-dimensional/multi-resolution data where the data are from multiple instruments.*
  - *c) Effective likelihood reconstruction methods and new graphical processing approaches.*

- ✓ Workflow architectures for Peta-scale datasets on distributed computing infrastructures.
  - *a) orchestration of compute intensive analysis of petascale datasets on distributed computing infrastructures (workflow engines on distributed systems, AAA protocols.)*

European data centres



Sketch of the current data flow from experiment sites to European data centers

Developments of new technologies followed by the scientific community require agreement and cooperation with research data centers that support the Astronomy and Astroparticle projects.

OBELICS strategy

- OBELICS will rely on existing e-initiatives:

  i) technical engagement with major computing data centres supporting the ESFRI projects such as those of the EU-T0 consortium.

  ii) survey available software products and services within major European e-infrastructures, such as EGI, PRACE and EUDAT.

- OBELICS will support cooperation with industry for innovation.

- OBELICS will organize training sessions for scientists to face the new challenges in computing and scientific software;

Conclusions

- Big-data challenges in Astronomy and Astroparticle Physics can be addressed through cross-fertilisation and shared development approach.
- The multi-wavelength scientific analyses and data-interoperability required by researchers imply important common developments.
- Synergies with Data Centres are needed in support of the implementation of ESFRI projects.
- Training initiatives for scientists involved in these challenges will help to adopt/explore new solutions.

➢ The ASTERICS H2020 project is the framework proposed by the A& A community.