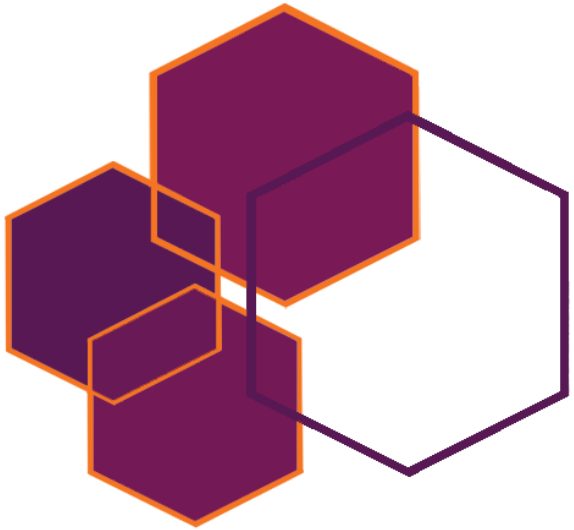


EU-T0

Giovanni Lamanna

LAPP - Laboratoire d'Annecy-le-Vieux de Physique des Particules,
Université de Savoie, CNRS/IN2P3, Annecy-le-Vieux, France

ERF, Big data & Open data

Brussels, 7-8 May 2014

cnrs IN2P3
The two infinities

EU-T0, Data Research and Innovation Hub

- **EU-T0** is a hub of knowledge and expertise that optimises the investment of the funding agencies in proven e-infrastructure by broadening, simplifying, and harmonising access, driven by well-defined user requirements.

- **EU-T0** aims to build a federated virtual European **Tier 0** data-management and computing centre, implementing the connection and coordination among the major national e-infrastructures.

- **EU-T0** contributes to the implementation of the vision towards a "general purpose European computing and data e-infrastructure for research"

[more agencies are joining]

Visions

- The e-Infrastructure Reflection Group (e-IRG) 2013 White Paper addresses the need of new paradigm for e-Infrastructure for research, able to account for **users needs** and clearly define **roles for all stakeholders**.

- CERN and the EIROforum members have published a vision for the evolution of the European e-Infrastructure aiming to create a **sustainable** IT environment **open to all science communities**. The vision capitalizes on the investments in computing infrastructure made over the last decade, and the facilities in place to support the major European Research Infrastructures (RIs).

- The next generation of services for the **distributed computing and storage infrastructures** have to address the current limitations and profit at best of the important advancement in **Cloud Computing** and in the **CPU architectures**. The next goal is to make available to Europe the next state of the art **distributed infrastructure for Big Data sharing and analysis**.

Rationale

- Considerable overlap between the research funding agencies in Particle, Nuclear, Astroparticle Physics, Cosmology and Astrophysics.

- APPEC promotes opportunities to cooperate in designing new software methods, computing and data processing infrastructures for the current and future major research projects.

- Promoting a large cluster of major research projects and key representatives around e-Science and data management issues.

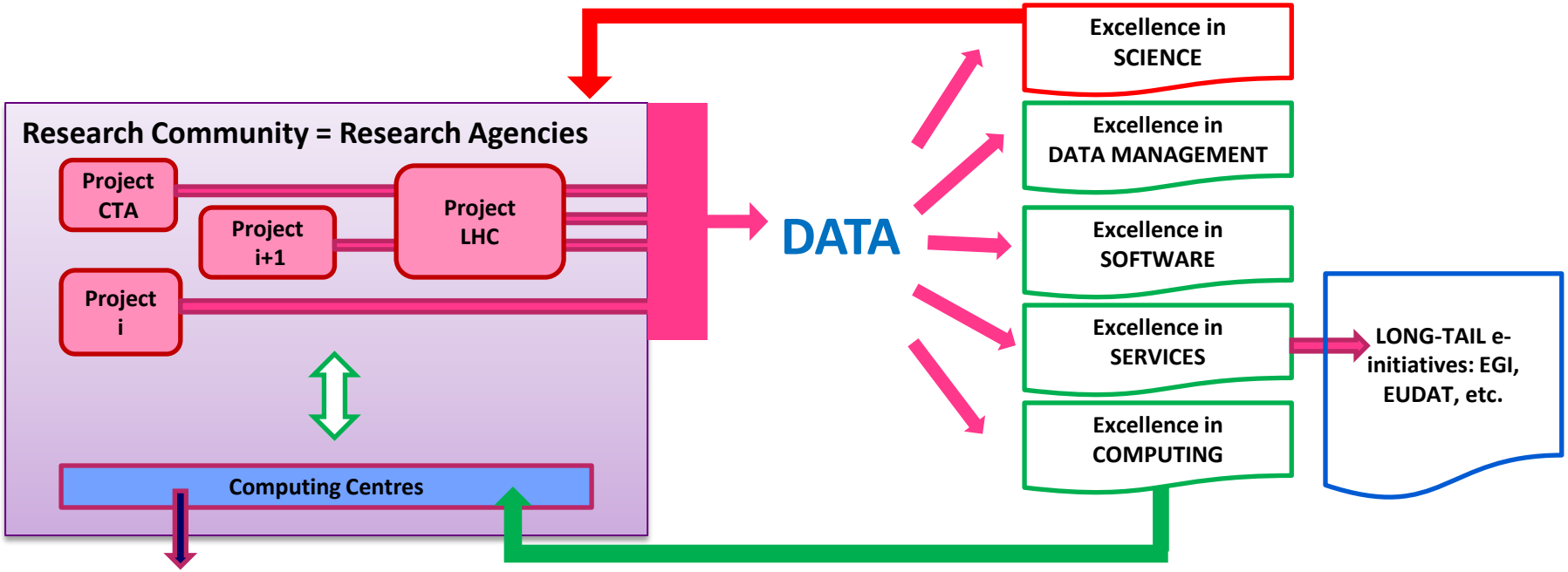
- The EU-T0 federation was launched officially on the 11th February 2014 by some major research institutes in the above mentioned domains: CERN, CIEMAT-ES, DESY-GE, IFAE-ES, IN2P3-FR, INFN-IT, KIT-GE and STFC-UK and the participation is going to be extended to more institutes.

- The approach aims to be a “core” project for a larger multi-domain federation.

Implementation

- Bringing the research communities closer each other to support their needs and: avoid fragmentation and repetitions; increase cross-fertilisation; share standards, expertise and developments; provide and share services; promote outstanding CCs in Europe.

- According to a "data and researchers centric" approach, the EU-T0 e-infrastructure accounts for the "user needs" and the expectations of research communities committed in major ESFRI RIs and ERF.



Implementation

E-INFRASTRUCTURES

A few pillar projects just started:

- ✓ The EU-T0 data backbone: heterogeneous storage managed in a federated way; interoperable (EUDAT-compatible); fulfilling the real-time ingestion and the archive access requirements;

- ✓ The EU-T0 cloud: extremely hybrid distributed computing architecture, scheduling, virtualizing and configurable for all users; serving also long-tail science.

- ✓ The EU-T0 VRE and software: organization, repositories and provision of services, preserving data, and new software programming test-bed provision across communities;

- ✓ The EU-T0 training: “data scientist” building profile and promoting careers.

- ✓ The EU-T0 pilot: new business model and interaction with private sector for co-developments around big-data services and cloud (Helixnebula).

Implementation

EU-T0 IN THE ECOSYSTEM

There is an important need for a number of coordination activities which embraces globally all e-infras. :

a) Organization and repositories for sharing services, software, data across communities;
b) Recognize shared/common needs for coordinated developments and services;
c) Efforts to develop new services/tools where needs are identified;
d) Policy development (shown to be essential for successful infrastructures);
e) Security coordination – policy development and incident handling;
f) Collaboration beyond Europe and with more scientific communities;
g) Integration with other forms of e-infrastructures including HPC and volunteer computing;
h) Coordination/negotiation with industry;
i) Creation of a training network to develop the key skills needed for the future including creating “data scientists”.

EU-T0: e-infrastructures background

- The EU-T0 partners are the research agencies owning large computing centers (CCIN2P3, CNAF, GridPP, PIC, DESY-Tier2, KIT-Tier1, … + many Tier2 national CCs) part of the Worldwide LHC Computing Grid (WLCG) project, having successfully implemented a distributed computing infrastructure but also …

- …Supporting large Research Infrastructures (RIs) (some in the ESFRI roadmap) in Astroparticle Physics and Cosmology, such as AMS, AUGER, H.E.S.S., MAGIC, CTA, FERMI, KM3Net, SKA, VIRGO/EGO and future gravitational waves projects, PLANCK, EUCLID, LSST, and in photon science XFEL, etc..

- The EU-T0 hub is built up on gathered resources currently of the order of hundreds of thousands of processing cores, targeting the half a million cores of computing resources in the next few years. EU-T0 archives big, heterogeneous and complex data through storage resources of the order of some hundreds Petabytes, which will grow up at the Exabyte scale already in the next years.

Particle Physics

- High capacity network is changing the participation of CCs and is impacting the LHC experiments computing models.

- The LHC experiments data throughput is already doubling in 2015 and the resources needs will increase further in the next years (millions of CPUs and hundreds PBs storage.)

- In order to exploit technological improvements we need to evolve the code and framework at least to:
  - fully use many-cores devices
  - move “big data” efficiently in the cloud paradigm
  - solve the high speed disk access

- The data “sharing”: preservation and long-term open access are new issues in the domain (see DPHEP talk)

Astroparticle Physics

- New RIs (CTA, EUCLID, LSST, SKA, GWs ...) are data intensive and have to deal with Big Data management issues:
  - pipelines from remote sensors at tens of GB/s rates;
  - on-site quick pre-processing and scientific alerts management;
  - huge archive and important data discovery requirements (hundreds PB/year);

  and Open Access issues:
  - high-level data products dissemination, curation and mining addressed through open access software, tools and services and the coherent framework of the Virtual Observatory (see BD in Astronomy talk).

Astroparticle Physics

- The Observatory functioning-mode (including limited data proprietary periods) and the multi-wavelength scientific analyses required by researchers implies important common developments:

  - Advanced, high speed, high capacity international networks among CCs archiving and processing data.

  - A functional Authorisation, Authentication and Accounting (AAA) infrastructure.

  - Research projects requiring access to all of the data at once will also require significant computational resources to achieve their goals -> cloud services and new workflow management on distributed compute-intensive systems.

  - Efficient Petascale DB queries for data analysis are critical-> A common work on the design and testing of extremely large database

Research Infrastructures

A few examples of projects and issues:

- CTA:
  - The pipelines will rely on a few CCs archiving O(100) PB of raw-data and need to build an efficient metadata model for data extraction and user access.
  - A worldwide community dealing with proprietary data (privileged users) and observatory products (guest observers). AAI is needed.
- EUCLID
  - Implementing a computing model where data archives and data processing are distributed among national CCs producing ~500 PB data from hundreds of TB raw and with a central metadata system.
- LSST
  - 15 TB/night during 10 years; final image collection : 0.5 Exabytes; ~10 million alerts on transient event per night; catalog of ~37 billions objects.
  - Extremely Large Databases, building Petascale systems (i.e. qserv) which is a highly scalable distributed database system to handle astrophysics queries.
  - Design a coherent approach to run cross queries between them.
- Gravitational Waves
  - Willing to create an international network of CCs for sharing data from different antennas and combined observations.
  - Long term preservation policies for important software tools, to read, manage and analyze the data (used for important discovery).

[…]

Conclusions

EU-T0 is a new path towards a hub of knowledge and expertise that optimises the investment of the funding agencies in proven e-infrastructures and driven by well-defined user requirements (including those concerning Big Data and Open Access).

Resources are a combination of hardware, software and skills.

Data management and data challenges are part of our ESFRI RIs and ERFs.

The excellence in science is based on excellent data produced by excellent facilities and managed by excellent e-infrastructures. They deserve support in Europe for frontier developments of potential transversal application.

++