

# EU-T0 Data Backbone

Tommaso Boccali,  
Luca dell'Agnello,  
Giacinto Donvito



# EU-T0 (1)



- Collaboration of major European research institutes and funding agencies
  - CERN, CIEMAT-ES, DESY-GE, IFAE-ES, IN2P3-FR, INFN-IT, KIT-GE and STFC-UK
  - Particle, nuclear, astro-particle physics to cosmology, astrophysics and photon science
- Aim to create a hub of knowledge and expertise in information technology and e-science
  - optimization of investment of the funding agencies in e-infrastructure



# EU-T0 (2)



- Aim to create a virtual Tier 0 center federating main European data management and computing centers
  - provision of software services and tools to the research communities
  - development of modern data management (and data preservation services and solutions)
  - deployment and operation of the federated computing infrastructure

# Data Backbone (1)

- Proposal for a project to build an infrastructure for Data Management ("Data Backbone") under consideration
- Room for an initiative complementary to EUDAT
  - targeting different communities/requirements
  - EUDAT oriented to many small research groups with small amount of data
    - E.g. LTDP Aleph use case (aggregate of  $\sim 30$  TB)
  - HEP collaborations cope with large amount of data ( $O(100)$  PB) at high rate ( $O(10)$  GB/s)

# Data Backbone (2)

- Goal: Develop an integrated pan-European scientific data warehouse with a backbone topology,
- Target communities: Astro-particle Physics, Astronomy, Cosmology, Nuclear Physics, High Energy Physics and Photon Science
  - but open to other disciplines with similar needs
  - Building on our experience in WLCG for new scientific collaborations (e.g. CTA)
- Functionalities of DM frameworks to become features of Data Backbone infrastructure
  - develop a set of tools for small collaborations
- Black-box approach for the user coupled with complete infrastructure elasticity
  - E.g. Site unavailability “not seen” by the user

# (Draft) Requirements (1)

- Interoperability with (at least) EUDAT
  - e.g. moving files from one to the other should be easily possible
- Guarantee custody and bit preservation of scientific data and corresponding metadata
  - Possibility to specify characteristics of data in the repository (e.g. custodial, temporary, scratch, access performance)
  - Recovery of replicas managed by the infrastructure (e.g. self healing of corrupted files)
  - Possibility to specify # and location of replicas
  - But also allow user “unawareness” of the actual data location

# (Draft) Requirements (2)

- Technology agnostic
  - heterogeneous storage resources managed by cooperative but independent administrative domains
  - Possibility to access via standard services and protocols (e.g. HTTP/WebDAV, posix-like access, object storage)
    - using both API from command line and graphic access from browsers;

# (Draft) Requirements (3)

- Able to cope with the (real-time) ingestion of huge amounts of (raw and derived) data
  - Including automatic addition of meta-data
  - Already tested at the scale of the LHC experiments (around 12 GByte/s).
- Unique Identifier for each file/object uploaded to the backbone

# (Draft) Requirements (4)

- Possibility to perform structured queries on the data (search for user-defined tags)
  - System must provide a metadata catalog programmable by the users
  - User defined labels to be associated with the data
  - System managed catalog with the association "logical file name" Data replicas

# (Draft) Requirements (5)

- Using standard federated identity and access right management solutions
  - E.g. EduGain
  - (Again) Close collaboration with EUDAT
- Data accessibility from any site
  - ACLs at single user level must be unique on the whole structure
- High speed network connectivity
- Computing capacity necessary to enable the data processing and a simulation environment integrated into the data backbone

# Other issues

- Collaboration with data owners to enable content preservation and ensuring future usability of data (including implementation of data life-cycle policies)
- Collaboration with consumer users in order to optimize the I/O performance of their algorithms

# Status & Conclusions

- Interest from PIC, INFN, IN2P3, CERN
- Phase of (advanced) brain-storming
  - First draft document under discussion
- Aim to be complementary to cloud project
  - But this infrastructure could also be used for backup, disaster recovery, data preservation etc...
- Need to converge in next few weeks
  - September call....