WP4 Thematic Service LAGO

A. J. Rubio-Monetro (CIEMAT)
on behalf of LAGO Collaboration

28.01.2020

Index

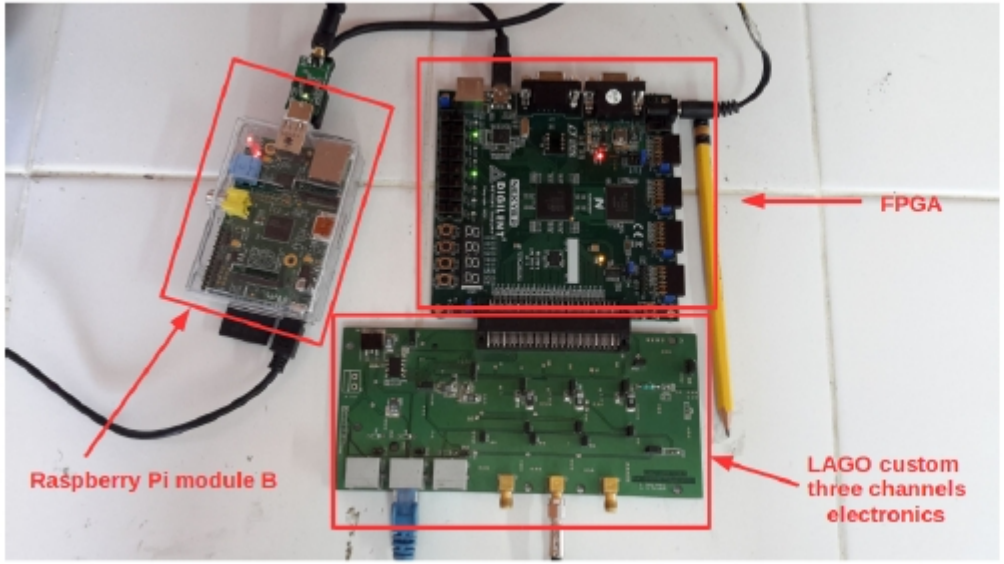
- Use Case Description
  - User community
  - Usage model
  - Workload characterization
  - Gaps and Bottlenecks
- Technical Description
- Expected Results
- Relevant Metrics
- Development Timeline

Use Case Description: User Community

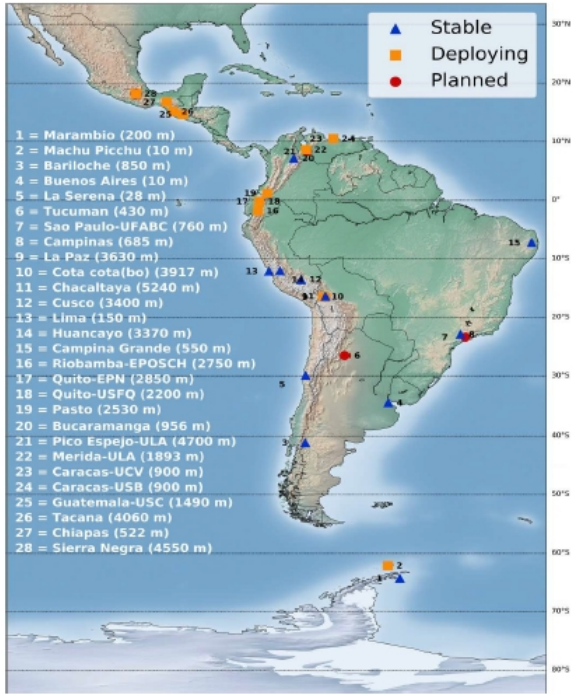
LAGO is an extended cosmic ray observatory composed of:

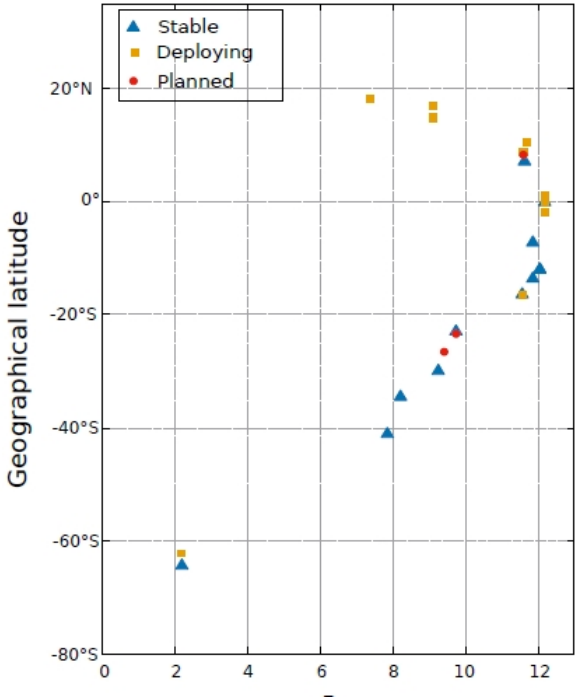
1) a network of Water-Cherenkov detectors (WCDs)
   - across Latin America & Antarctica
   - different altitudes and latitudes
   - 4 in production 24/7, 11 installed, 10 planned.
2) a consortium
   - 99 researchers
   - 11 countries
   - 29 institutions

Use Case Description: User Community



Use Case Description: User Community

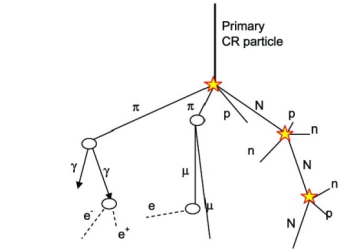




Use Case Description: Usage model

- Certain institutions performing the pipeline of real data: I/O of raw data, storing and analysis of intermediate and final results.
- Researchers also make use of HPC facilities for simulations
- CORSIKA (and GEANT4) mainly used.
- A CKAN repository at UIS (Colombia) gather all the data.
- Data-sets contain:
  - Metadata in DublinCore format
  - 1 hour of measurement or simulation
  - 1 file (measurements) or 60files (1 per minute in simulations)

Use Case Description: Usage model



Current pipeline for the measured data.

LAGO sites

LAGO CPU and storage sites

Preprocessing (Correction by the atmospheric pressure)

LAGO CPU and storage sites

Processing (Sites and events characterization)

LAGO repositories

Histograms
Papers

Data level: L0
Raw data

Data level: L1
Low resolution

Data level: L2
Corrected scalers

Data level: L3
Research output

Use Case Description: Usage model

Data Type	Source	Owner	Visibility
Raw (L0)	Water-Cherenkov detector (WCD)	LAGO	Private while analysed data are not available (public).
Cleaned (L1)	Raw data from WCD		
Analysed (L2 and L3)	Cleaned data from WCD		Public after fixed waiting period
Simulated	1 User	User	Public after variable waiting period

Use Case Description: Workload Characterization

CPU	- 1 pre-processing or analysis job (3,600s) = 0.6 CPU hours - 1 simulation job (60s) = 0.25 - 10 CPU hours **Total = 4,248.9 CPU hours/week (~ 26 cores/day)**[1]
Numb. of jobs	**Total = 2,366 jobs/week (338 jobs/day)**[1]
RAM 1 job.	4GB (maximum)
Storage 1 job.	**Usually: 10GB**, rarely 100GB (event simulations)
How the application is run?	- Real data: single large experiment associated to each detector. - Simulations: variable demand, ~ 1,500 jobs/month[1]
Input data 1 job	- Pre-processing and analysis : ~ 100 -200 MB (one file compressed) - Simulations: order of KB (one file compressed)
Output data 1 job	**Usually: 100-200MB,** rarely 2GB.

[1][4 WCDs + 25 users (1 sim./month)]

Use Case Description: Storage Characterization

Permanent storage[1]	5.6 - 11,6 TB/year
Granularity[1]	**157,680 files/year (in 103,980 reachable data-sets)** 50KB - 2GB files (**usually 100-200MB** compressed files)
Bandwidth required	Regular GEANT one is enough.
Access pattern	All: private non curated, privated curated, public...
Does the application need to be accessed from outside (e.g. web application)?	Could be a possibility to better integration with EOSC
Does the application require downloading data from external sources? Yes	Yes from LAGO data repository (but the idea is to directly store WCD raw data in EGI DataHub previously to computation)

[1][4 WCDs + 25 users (1 sim./month)]

Use Case Description: Gaps and Bottlenecks

- Users work with local/remote HPC facilities:
  - non homogeneous
  - do not cover demand peaks
  - without centralised accounting
- LAGO has been storing data at UIS (Co) with CKAN and performing some simulations at CIEMAT (Es), but:
  - there is no option for reproducibility activities
  - many users do not upload simulations to repository

Technical Description

Service Scope	Service used now	Service planned	Limitation
AAI	User/password	EGI Check-in.	- B2 & DataHub (mainly) should work fine on private clusters (compatibility). - Data confidentiality before waiting period. - Estimated: 5.6-11.6 TB/year
Workload Mng.	Cluster batch	Cluster batch & (EC3 or EGI Workload Service)	
Resource Mng.	Cluster batch	Cluster batch & IM+EC3/VMops	
Data Storage & Storage Resources	Local filesystems & UIS repository	B2FIND, B2HANDLE, EGI DataHub	
Monitoring	N/A	ARGO	Some accounting and monitoring is needed for continuous processing raw to analysis
Computing Resources	Local HPC clusters	HPC clusters & FecCloud	To support the continuous processing of real data: 12 cores/day

Technical Description Justification

- EGI DataHub (OneData):
  - It allows categorise L0-3 data and permissions based on VO roles.
  - Metadata can be automatically gathered by other services such (B2FIND) and adding PIDs (B2HANDLE)
  - Theoretically, it can be accessed from different Comp. infrastructures (e.g. local clusters) using EGI AAI through the OneClient docker image.
- Modified Oneclient docker image to be distributed on:
  - Pre-deployed FedCloud VMs or cluster nodes
  - CORSIKA release and scripts to validate and to store/publish output files in DataHub
- No preference tools for deploying VMs and manage production:
  - Coud be IM/EC3, VMOps, EGI WS (DIRAC), INDIGO, or combined

Expected Results

- Computing with EOSC services:
  - Properly process data of WCDs in production
  - Facilitate the simulation process to users
- Generate and store data following FAIR guidelines to
  - Maintain and expand the collaboration through time
  - Correct processed data with new software
  - Properly re-use simulations (without waste CPU time)
- Increase the LAGO impact in the iCosmic Rays community and other disciplines

Relevant Metrics

User Metrics	
1	MU_NUS
1	MU_NUSA
2	MU_NIU
2	MU_NIUA
1	MU_NCEA
1	MU_NCOA
1	NRUSA

Cap. Metrics	
3	MC_NSE
3	MC_NSEA
4	MC_CPU
4	MC_CPUA
4	MC_MEM
4	MC_MEMA
4	MC_STO
4	MC_STOA
4	MC_MXCC
N/A	MC_MXCCP
4	MC_MXMCP
4	MC_MXTHR

Sci Impact Metrics	
5	MO_PUB
5	MO_COM
6	MO_TRAH

Synergies Metrics	
8	MF_COSH
9	MF_JDIS
10	MF_SYN

Usability Metrics	
7	MU_PER
7	MU_ERR
7	MU_SCA
7	MU_COM
7	MU_INT
7	MU_LC
7	MU_CON
7	MU_ROB
7	MU_OVA

1) VO+group / LAGO service
2) Accounting: registered access to restricted data & anonymous access to public data (B2FIND+DataHub)
3) LAGO Service
4) Accounting (EGI)
5) LAGO Publication Repository
6) WP6
7) Computer System Usability Questionnaire
8) Num. of downloads (AppDB)
9) WP1
10) Annotation / Manually

Development Timeline