

HEPiX

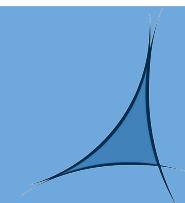


# PIC Report - J. Flix

[on behalf of PIC team]

HEPiX Autumn-Fall 2020 / Virtual

12-16 October 2020



PIC  
port d'informació  
científica



Institut de Física  
d'Altes Energies



Ciemat

Centro de Investigaciones  
Energéticas, Medioambientales  
y Tecnológicas



# PIC in numbers

October 2020

CPU: 110 kHS06  
Disk: 10.1 PB  
Tape: 31.9 PB



**Spanish WLCG Tier-1 centre** → ~80% of resources

→ Provides ~5% of Tier1 data processing of CERN's LHC detectors ATLAS, CMS and LHCb

**¼ of the Spanish ATLAS Tier-2** and **a Tier-3 ATLAS data analysis facility** → ~10% of resources

**T2K** [neutrinos], **MAGIC** and **CTA** [gamma-ray astronomy], **PAU** and **EUCLID** [cosmology], **VIP** [instrumentation], opportunistic access to **LIGO/VIRGO** and **DUNE**, among others...

# PIC farm updates

322 compute nodes (8092 slots), under **HTCondor v.8.8.10**

→ 84% of compute nodes dual-stack; old hardware still in IPv4 (won't be migrated)

**2x HTCondor-CE v3.4.2-1**

**2x ARC-CE v.6.6.0 (used by ATLAS as HPC gateways - see later)**

**10 GPUs available:** 2 in use for farm jobs (VIRGO/LIGO users) and 8 for JupyterHub

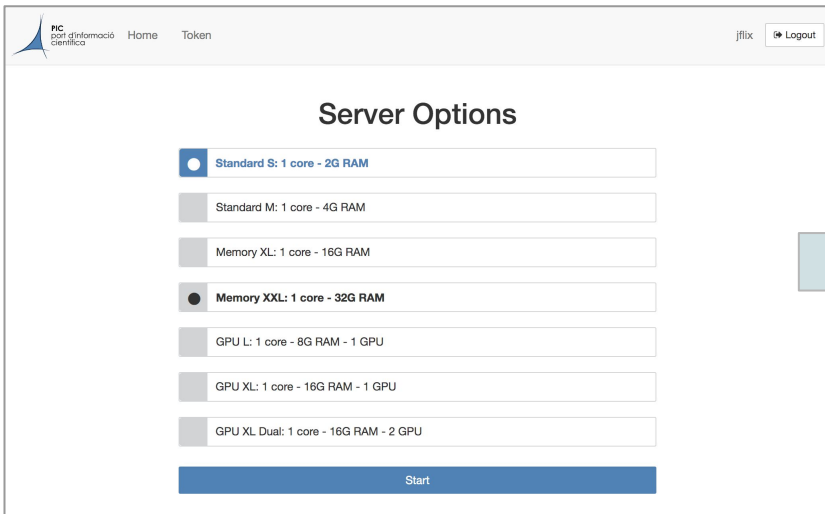
Continued tests at low scale with CMS workloads on **AWS** and **CloudSigma**

A portal is available to select different profiles (CPU, mem, GPU) to **spawn Jupyter Notebooks to the PIC farm** ➔

# PIC farm updates

Web browser  
(<https://jupyter.pic.es>)

Web browser  
JupyterLab session



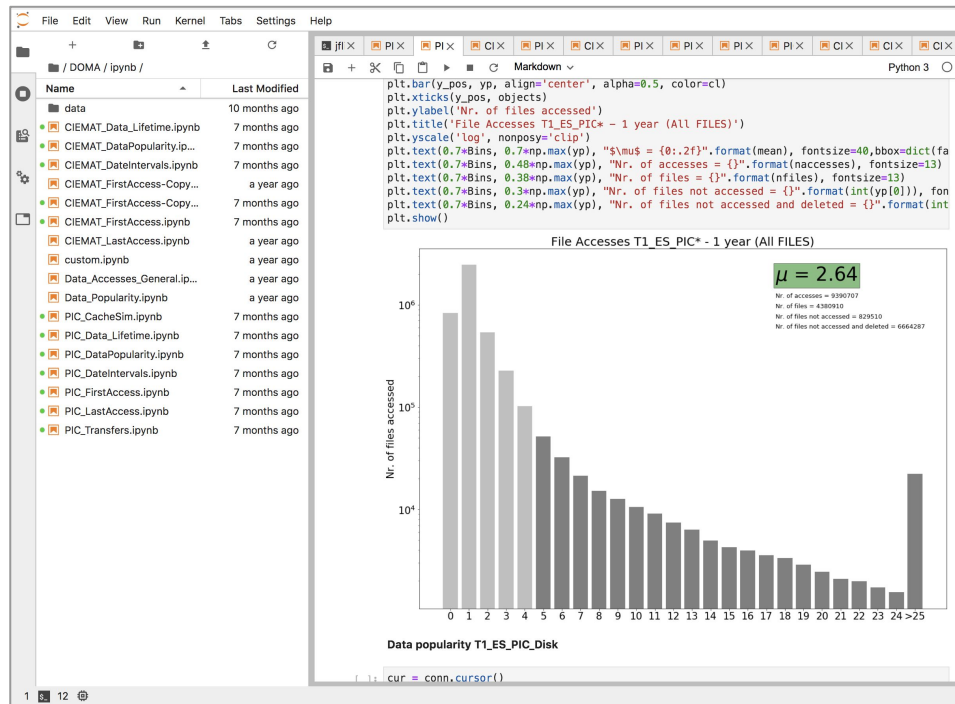
PIC port d'informació científica Home Token jflix Logout

## Server Options

- Standard S: 1 core - 2G RAM
- Standard M: 1 core - 4G RAM
- Memory XL: 1 core - 16G RAM
- Memory XXL: 1 core - 32G RAM
- GPU L: 1 core - 8G RAM - 1 GPU
- GPU XL: 1 core - 16G RAM - 1 GPU
- GPU XL Dual: 1 core - 16G RAM - 2 GPU

Start

Users can then open a unique JupyterLab session in the running slot, and can execute multiple notebooks



File Edit View Run Kernel Tabs Settings Help

/ DOMA / ipynb /

Name	Last Modified
data	10 months ago
CIEMAT_Data_Lifetime.ipynb	7 months ago
CIEMAT_DataPopularity.ipynb	7 months ago
CIEMAT_DateIntervals.ipynb	7 months ago
CIEMAT_FirstAccess-Copy...	a year ago
CIEMAT_FirstAccess.ipynb	7 months ago
CIEMAT_LastAccess.ipynb	a year ago
custom.ipynb	a year ago
Data_Accesses_General.ip...	a year ago
Data_Popularity.ipynb	a year ago
PIC_CacheSim.ipynb	7 months ago
PIC_Data_Lifetime.ipynb	7 months ago
PIC_DataPopularity.ipynb	7 months ago
PIC_DateIntervals.ipynb	7 months ago
PIC_FirstAccess.ipynb	7 months ago
PIC_LastAccess.ipynb	7 months ago
PIC_Transfers.ipynb	7 months ago

```
plt.bar(y_pos, yp, align='center', alpha=0.5, color=c1)
plt.xticks(y_pos, objects)
plt.ylabel('Nr. of files accessed')
plt.title('File Accesses T1_ES_PIC* - 1 year (ALL FILES)')
plt.yscale('log', nonposy='clip')
plt.text(0.7*Bins, 0.7*np.max(yp), "$\mu$ = {:.2f}".format(mean), fontsize=40, bbox=dict(fa
plt.text(0.7*Bins, 0.48*np.max(yp), "Nr. of accesses = {}".format(naccesses), fontsize=13)
plt.text(0.7*Bins, 0.38*np.max(yp), "Nr. of files = {}".format(nfiles), fontsize=13)
plt.text(0.7*Bins, 0.3*np.max(yp), "Nr. of files not accessed = {}".format(int(yp[0])), fon
plt.text(0.7*Bins, 0.24*np.max(yp), "Nr. of files not accessed and deleted = {}".format(int
plt.show()
```

File Accesses T1\_ES\_PIC\* - 1 year (ALL FILES)

$\mu = 2.64$

Nr. of accesses = 939707  
Nr. of files = 4389910  
Nr. of files not accessed = 829510  
Nr. of files not accessed and deleted = 6664287

Nr. of files accessed

0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 >25

Data popularity T1\_ES\_PIC\_Disk

```
cur = conn.cursor()
```

Terminal session for the users. They can use their own Python environments

# HPC resources exploitation by LHC [Spain]

Collaboration agreement between **Barcelona Supercomputing Center (BSC)** and **LHC Computing Spain** to exploit a fraction of their resources for ATLAS, CMS and LHCb (\*).

→ <https://www.bsc.es/marenostrum/marenostrum/technical-information>

(\*) LHC computing designated as one of the BSC strategic projects

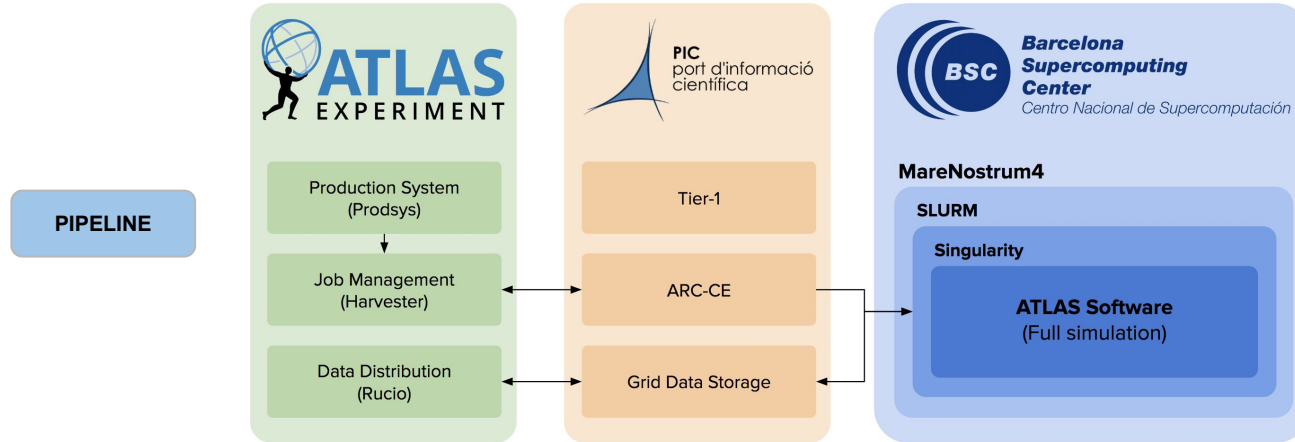
BSC will host one of the first **pre-exascale supercomputers in EU**: ~200 peak Petaflops

BSC has **some limitations** to run WLCG jobs:

- Execute nodes do not have internet connectivity, hence it breaks late binding models used in WLCG
- Not possible to install edge services (Squids for conditions and CVMFS, ...)

**ATLAS** and **CMS** have opted for different solutions to overcome the lack of Internet connectivity from the execute nodes @ BSC ➔

# Use of the BSC by ATLAS PIC Tier-1



Using two **ARC-CEs** at PIC to interconnect MareNostrum and ATLAS production system

Only simulation workflow validated - singularity containers, pre-placed at MareNostrum GPFs

MareNostrum accepts only SSH protocol for job submission and data transfer

# Use of the BSC by ATLAS PIC Tier-1

Submitting **ATLAS** payloads to BSC from PIC Tier-1 since 2018, in production since 2019

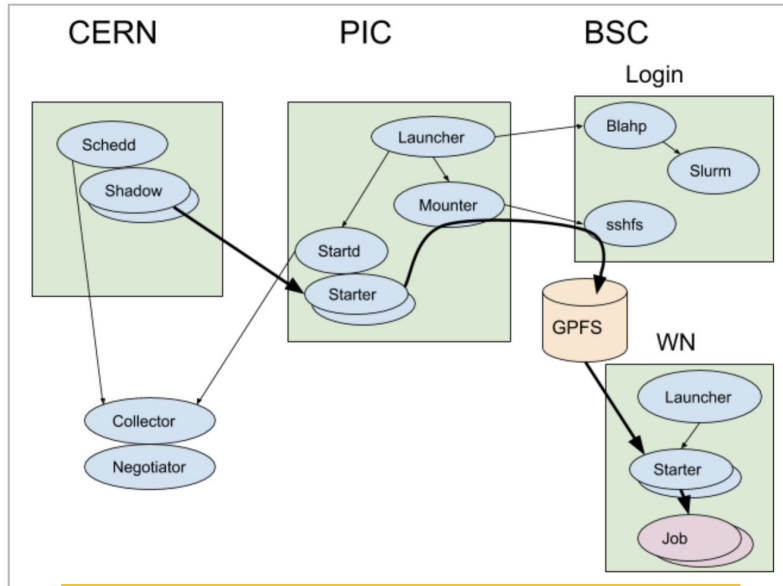
**Stats for 2020:** 6 million hours approved in 2020 (83% used) and one request of 4 million hours pending approval [1 hour = 16.75 HS06-hours] → 53 millions of events simulated



In addition, ATLAS Tier-2s in Spain send jobs to MareNostrum. In 2020, 15% of CPU has been obtained from HPC resources.

# Use of the BSC by CMS PIC Tier-1

PIC and HTCondor team collaboration to use a shared FS as control path for HTCondor



Setup that interconnects all of the HTCondor daemons for the CMS Global Pool, PIC Tier-1 center and the BSC

Allocation of 1Mhours @ BSC for CMS submitted [Oct-Dec. 2020]

## Current status

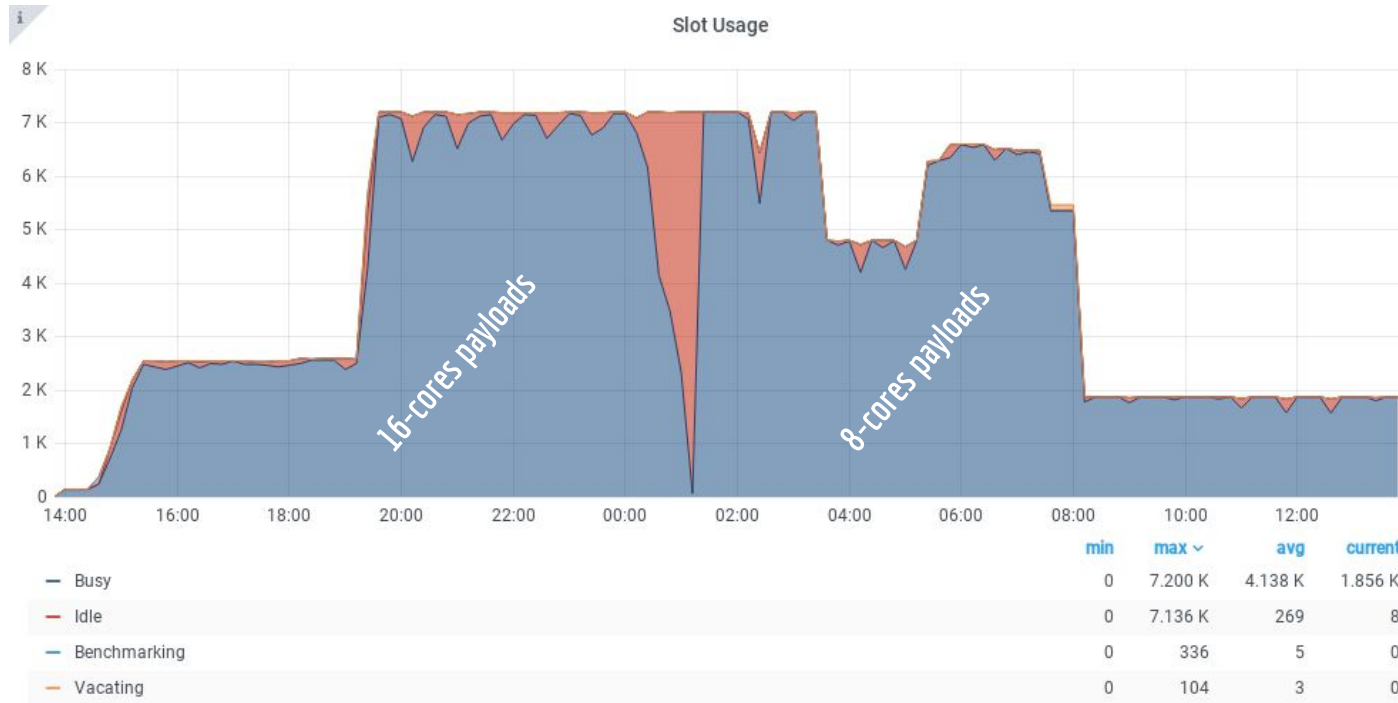
- An HTCondor-bridge has been deployed at PIC to interact with BSC execute nodes through the login node, mounting the shared FS through **sshfs** and sending jobs to the Slurm scheduler via **ssh**
- Established a procedure in CMS to prepare self-contained payloads which do not require external connectivity (application packaged inside a **Singularity container**, and **conditions data** read at run time dumped into a **sql file**)
- Testing the HTCondor bridge to submit and run a CMS singularity image @BSC
- Integration tests running CMS pilots in the HTCondor bridge that connect to the CMS global pool
- Working on matchmaking policy to acquire specific jobs suited to BSC

## Next steps

- Build singularity images for BSC production purposes
- Developing a data management service to get output files from BSC



# Use of the BSC by CMS PIC Tier-1



**Scale tests:** running singularity images for CMS simulation on ~7k slots (running in 48 cores machines, tuning payload core usages to maximize global CPU efficiency), plugged into the CMS Global Pool (test instance) through PIC HTCondor infrastructure, using the shared FS at BSC

# New tape library

IBM TS4500



New **IBM TS4500** in production: 1 frame L55 + 6 LT08 drives  
→ ~5 PB capacity installed with cartridges LT07 M8

We have experienced positioning errors for LT08 drives  
→ Reported to IBM - several firmwares tried - v.M570 stable (since Jul.'20)

SL8500

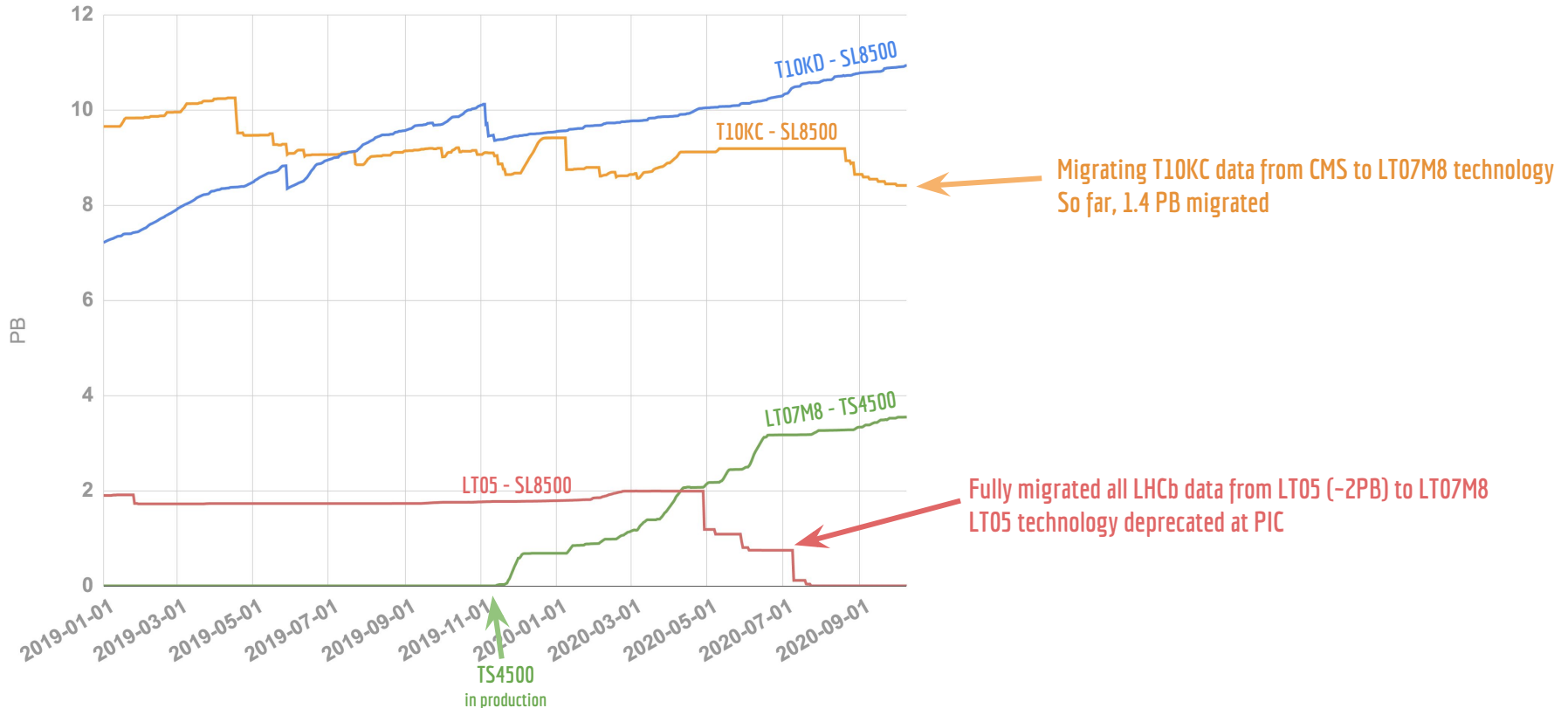


This new IBM library is expected to grow to host future data  
→ It will host new data and data migrated from SL8500 library  
→ Dedicated drives, frames and cartridges will be installed to handle this  
(writing the tender for a new D55 frame + 2 drives LT08)

PIC currently runs Enstore 6.3.4-2 (CentOS7)

# Data migrations to TS4500

## Used space by WLCG



# PIC storage updates

~10 PB running on **dCache 5.2.30**

- dCache pools in dual-stack
- TPC enabled for HTTPs and XRootD and token authentication (PIC in DOMA testbeds)

**StashCache** deployed on K8s (OSG repo) for Virgo/Ligo

- 3.2 TB - 85% occupancy
- Some issues are being addressed to migrate it to XRootD 5.0

**xCache** deployed (OSG repo) for the CMS experiment

- 4TBx36, 16 cores L5630 (HT enabled), 48 GB RAM, 10 Gbps - ~50% occupancy
- Currently at low scale. Monitors and proper setting being deployed and/or investigated

# PIC storage updates

Using **Ceph** with old HW (pools and switches) for non-critical services

→ backups, scratch areas, virtualization test instance

→ At the moment 1 cluster in production + 2 test clusters [\[ex: CephFS scratch space for Euclid project\]](#)



## Production cluster (Nautilus v. 14.2.11)

3 Monitor servers, 3 MDS servers, 2 iSCSI gateways and 7 OSD servers

Each OSD server has:

- 36 x 3TB SATA HDD + 1 x 2TB NVMe (WAL/DB)
- 168GB RAM
- 2 x 10Gbps (public network) + 2 x 10Gbps (cluster network)

\* 252 OSDs, 700TB raw capacity

\* Replicated pools and EC pools (k=4,m=2).

\* Using RBD and CephFS

## Test Clusters

\* Test Cluster 1 (Nautilus v.14.2.11) -> Testing RGW

\* Test Cluster 2 (Octopus v.15.2.5) -> Testing Octopus new features

# ESCAPE project: PIC data injector

PIC contributes to the ESCAPE project

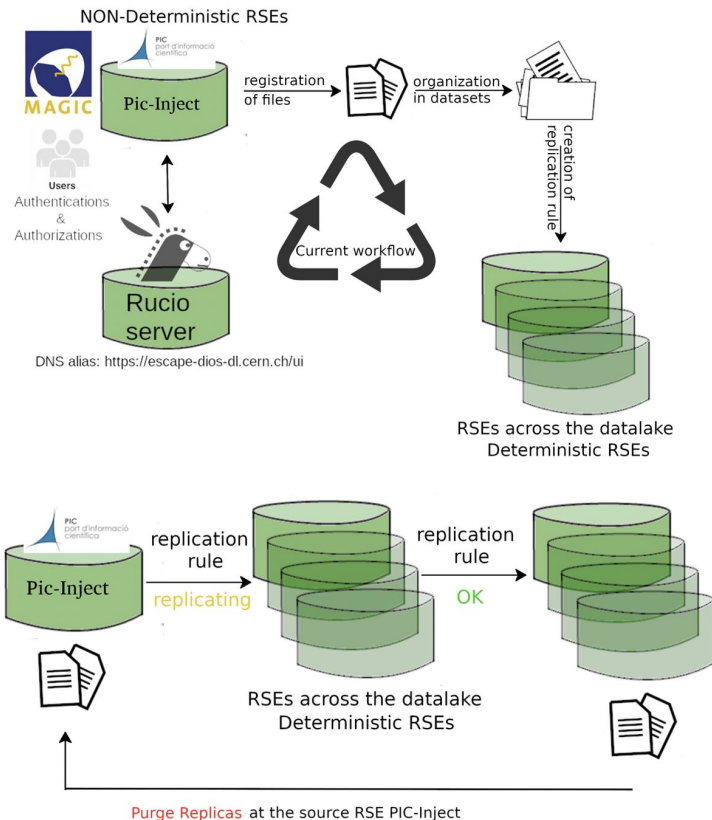
**Aim:** develop solutions to handle the large datasets produced by Gamma ray telescopes, adopting Rucio to stream files from the telescopes to a Data Lake for permanent storage and access

Using **PIC and Cherenkov telescopes in Canary islands** as the testbed (MAGIC and CTA)

## Currently testing

A Rucio SE (RSE) configuration at PIC that allows to register files with their original path at the detector (mimicking the future config of an RSE at origin)

Automation of the orchestration of the data using Rucio replication rules to different research facilities among the Data Lake. Once replicated, source files are removed





Thanks!  
Questions?

**Credits to:** E. Acción, V. Acin, C. Acosta, A. Bruzzese, J. Carretero, J. Casals, R. Cruz, M. Delfino, J. Delgado, J. Flix, G. Merino, C. Neissner, A. Pacheco, C. Pérez, A. Pérez-Calero, E. Planas, M.C. Porto, B. Rodríguez, P. Tallada, F. Torradeflot