PIC Report - J. Flix [on behalf of PIC team]

HEPiX Autumn-Fall 2019 / Amsterdam

14-18 October 2019

PIC
port d'informació
científica

Institut de Física d'Altes Energies

Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

PIC in numbers

October 2019

CPU: 110 kHS06
Disk: 9.6 PB
Tape: 30.4 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** [new] and **DUNE** [new]

(some) news at a Glance

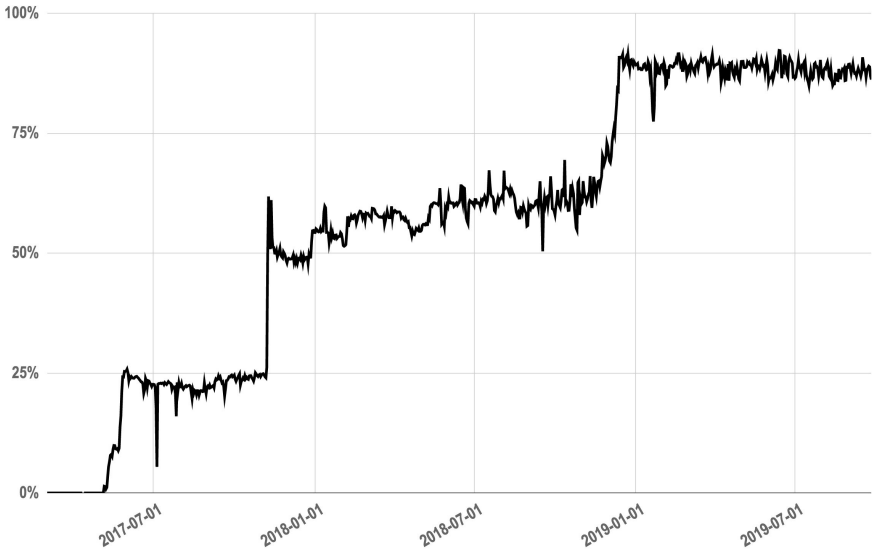
- ~90% of PIC farm managed by **HTCondor** **[Torque/Maui decommissioning this October '19]**
- **JupyterHub** cluster deployed in PIC, handled by K8s
- **GPUs**: offering GPUs for interactive **[Jupyter]** and batch **[HTCondor]** use
- **AWS** cloud bursting tests and first tests with **Glacier Deep Archive**
- **dCache** version 4.2.32 → migrating to **5.2** before the end of the year **[DOMA TPC compatibility]**
- New **Ceph** storage cluster deployed (~400 TB raw capacity)
  - CephFS scratch space for Euclid project
  - RBD+iSCSI as storage backend for virtualization test platform
- Recent **purchases**:
  - Tape library IBM TS4500 (1 frame), 4x LT08 drives, ~4 PB LT07 M8 tapes, ~1 PB disk
- Participation in **ESCAPE** and **ARCHIVER** EU projects
- Active participation in several **WLCG/HSF working groups**: DOMA Access, DOMA TPC, WLCG/HSF Systems Performance and Cost Model, ...
- **PAU survey** data center hosted at PIC, in charge of the **PAU data management** **[publication]**

HTCondor migration at PIC

PIC port d'informació científica

HEPiX

Fraction of WNs migrated to HTCondor



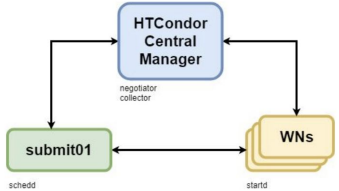
The migration of PIC to HTCondor is **almost done** with the exception of some small groups of users

1xGPU recently made available through HTCondor

Torque/Maui to be stopped by end of October

→ But, will keep temporary (and small) CREAM-CE/Torque-Maui for EGI ops

HTCondor v8.8.4 (stable version)
2 Central Managers in HA [CentOS7]
2 HTCondor-CEs Tier-1 (v3.2.1) [CentOS7]
1 HTCondor-CE Tier-2 (v3.2.1) [CentOS7]
WNs in **CentOS7**

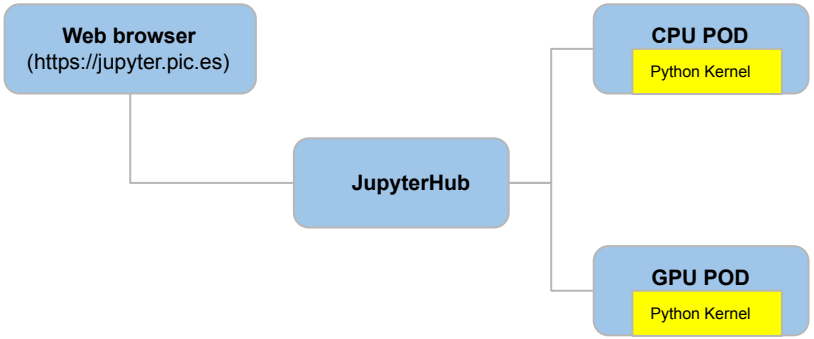


Local submissions schema

JupyterHub on Kubernetes

- Interactive sessions
- GPU / CPU support
- Jupyter Notebook Integration
- Docker Service

Basic schema



In progress:

- HTCondor integration
- Improve accounting system

- Functionality Jupyter Notebooks
  - Persistent Storage: **Ceph & NFS**
  - **LDAP**: user impersonation
  - Access to data, mount points, etc..
  - **GPU** configurations
- Monitoring & Logging
  - **Prometheus / Grafana**
  - **ELK**

Amazon - cloud bursting tests

We tested **AWS** for a week (June 2019), doubling PIC compute power

- Integration of a cloud environment with the local batch system - sporadic increase of resources
- Special interest in a spot instance based scenario

Data center in Frankfurt (**~40 ms**) - used Condor_Annex

Set up HTCondor Connection Brokering (CCB)

- **Bridge** server to connect the local system to the outside nodes

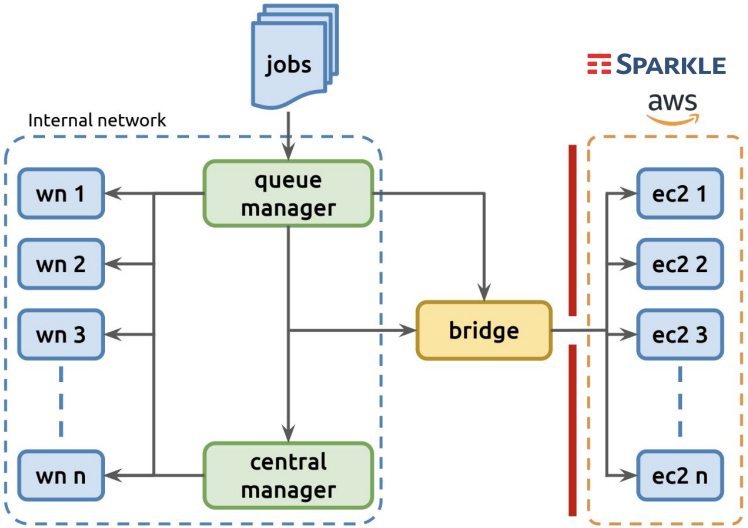
HTCondor-CE routing modified so only **ATLAS** and **CMS** send jobs to AWS

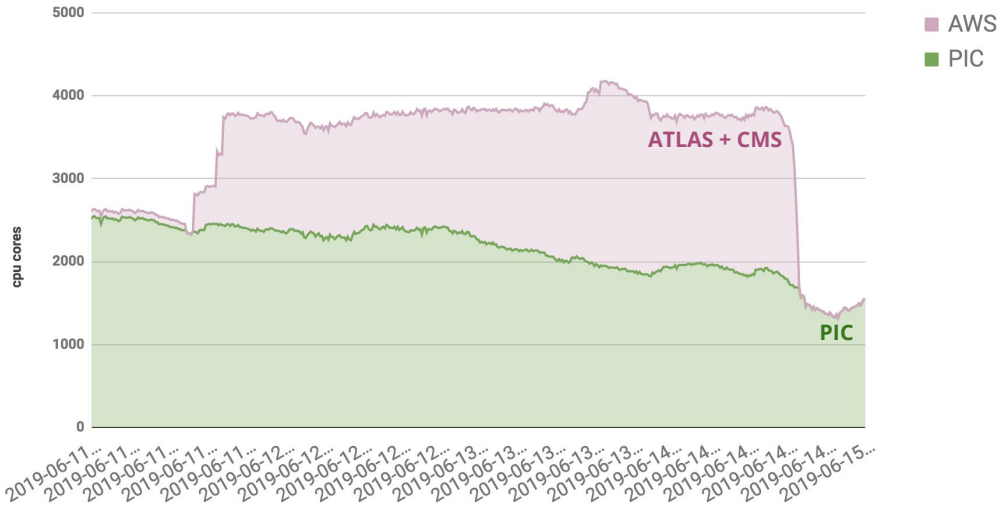
Custom **WN image** deployed in AWS servers, + CVMFS, + access to Squids

Configuration of **spot instances requirements** during the test

+info in this talk by J. Casals [IberGrid 2019 conference]

Amazon - cloud bursting tests



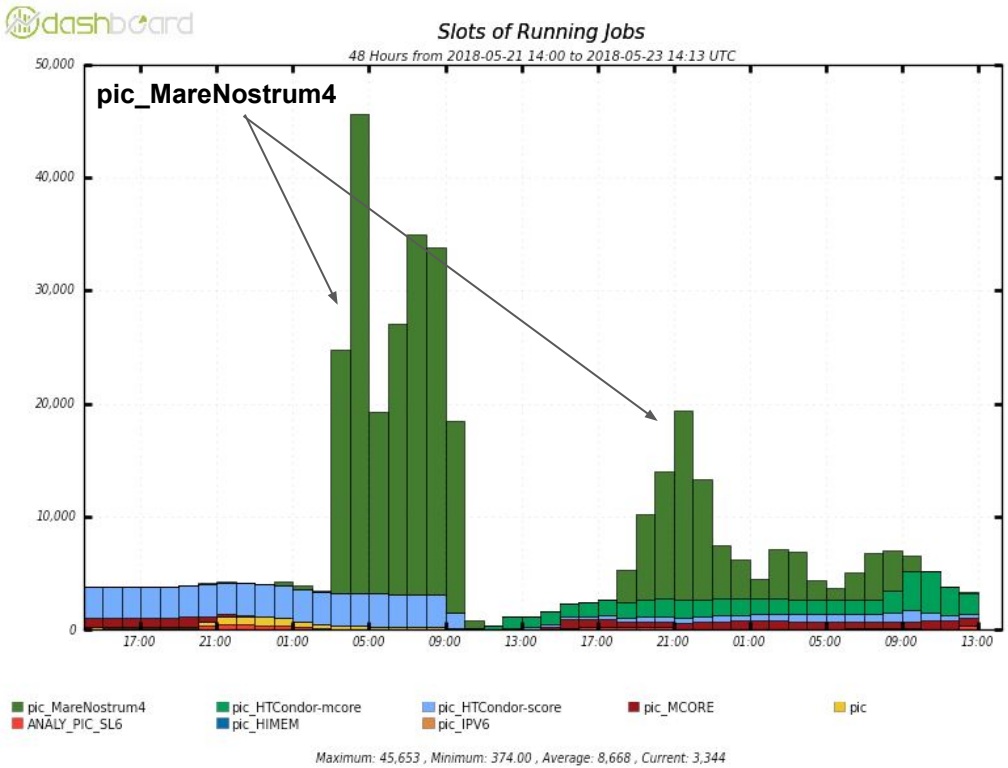


Good option to increase computing resources sporadically

Flexible and easy to deploy through HTCondor

Not very good for data intensive jobs **[see later in this talk]**

Integration of HPC resources [ATLAS]



Tests on the **MareNostrum HPC** integration in the ATLAS production system started in April 2018 in joint collaboration with IFIC Tier2 site

Since then, we have received hours to exploit Spanish HPC's (**RES** and **PRACE**):

In 2019, PIC has been granted **2.75 million hours in the MareNostrum 4 HPC**

Two types of payload submission:

- 1 job = 1 full node (48 cores)
- 1 job = 50 nodes using MPI/Yoda (2400 cores)

Data async. transferred to PIC and registered into ATLAS Rucio system

Tested transfer mode using **globus-url-copy** with ssh as authentication (no certificates) which is standard for HPC sites

Integration of HPC resources [CMS]

For CMS, we are **working in a model**, similar to ATLAS, in close collaboration with HTCondor developers

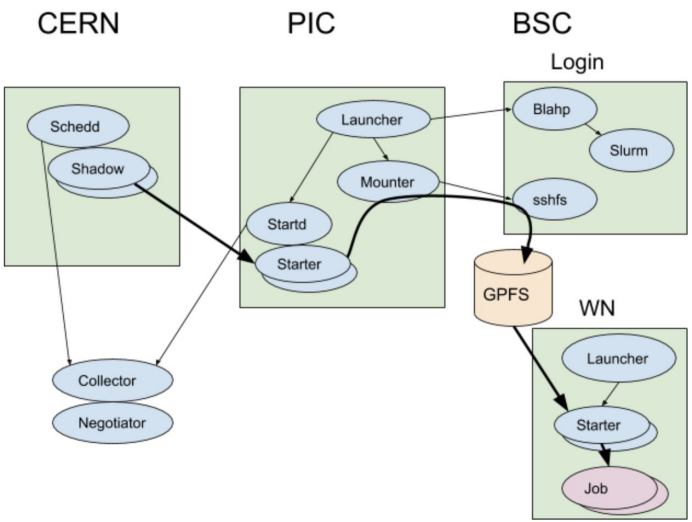
Lack of outbound network connectivity in nodes reduces flexibility for CMS… since CMS pilot jobs connect to global CMS HTCondor pool to get the actual payloads

→ Developed and tested a mechanism to interconnect HTCondor pools through **shared file systems**

→ Dedicated **PIC testbed** in place, some tests run connected to CMS Global Pool

→ We need to **instrument** the CMS **payloads** so they can run at BSC (sqlite file for conditions, singularity image, data export handling)

→ Goal is to incorporate BSC resources and run CMS simulations



HTCondor High Throughput Computing

What's new in HTCondor? [EU HTCondor Workshop 2019]

New tape library

IBM TS4500

NEW

New **IBM TS4500**, with 1 frame L55 and 4 LT08 drives
447 tape cartridges LT07 M8 (~4 PB) installed

Already installed and fully integrated in a **test instance**

→ By the end of October it will be added in production

SL8500

(some) CMS data will be then **migrated** from T10KC (SL8500) to LT07 M8 in the new IBM TS4500 library

This new IBM library is <u>expected to grow</u> 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

Network

PIC current WAN at **20 Gbps**

Activated NetFlow in our Nexus 7009 router and integrated in Elasticsearch **[next slide]**

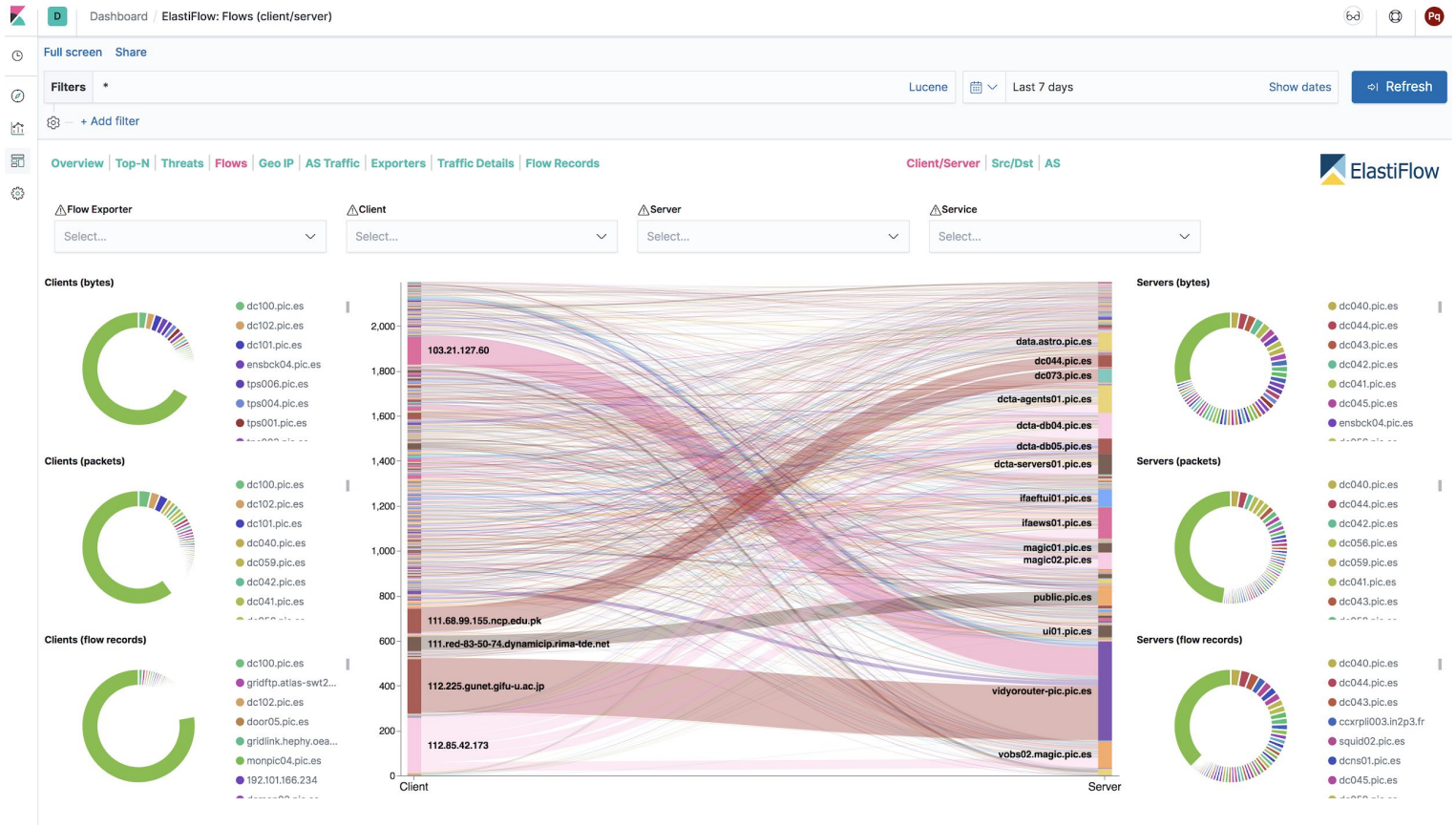
RedIRIS is tendering to enable 100 Gbps across Spain → **RedIRIS-Nova at 100 Gbps**

PIC would increase it WAN connectivity to 100 Gbps by mid-2020 - proposal submitted for funds to replace/buy all of the needed hardware

Proposal based on Leaf-Spine Network Topology

- 2x Spine → total of 64x 100 Gbps ports
- 6x Leaf → total of 288x 25 Gbps ports, and 8x uplinks of 100 Gbps

ElasticFlow



Towards a regional federation [CMS]

While ago CMS enabled **overflow** of analysis jobs from PIC to CIEMAT (Madrid) and vice versa, and we deployed a regional XRootD re-director in HA

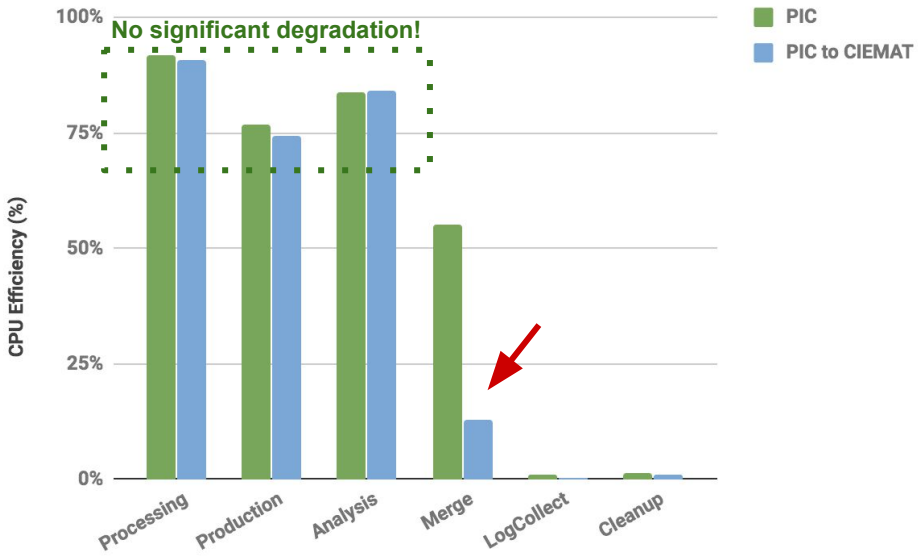
Since May 2019, we are **flocking** CMS pilot jobs from PIC to CIEMAT and vice versa, since we have HTCondor BS in both sites → 80 cpu-cores available at each site **[dedicated machines, for the moment - 10 ms latency]**

Regional input file reads are preserved, since we have regional XRootD re-director deployed - hence we can study job degradations when running remotely

How does latency affect the CMS workloads? This is important to understand the effects of federating the resources at a national level

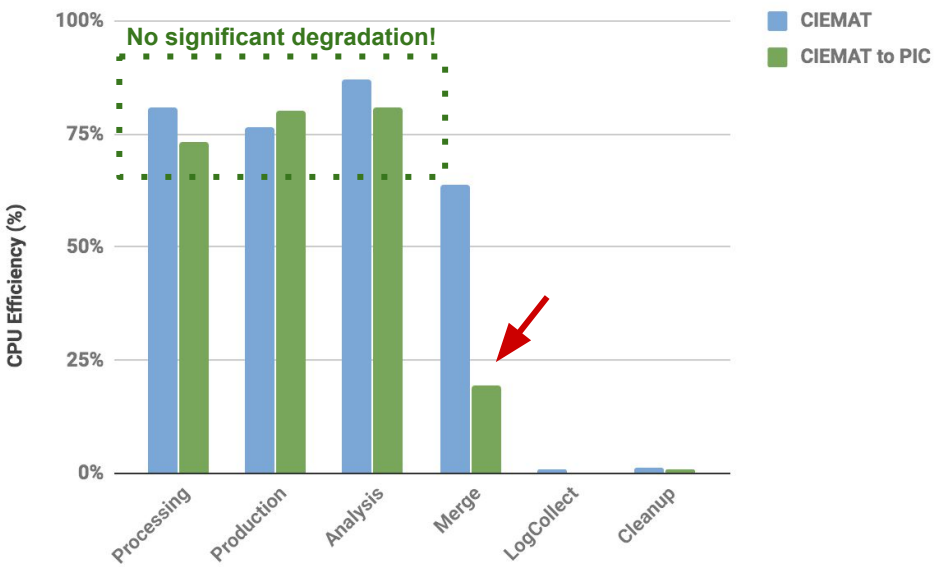
Towards a regional federation [CMS]

From 2019-06-07 to 2019-07-07



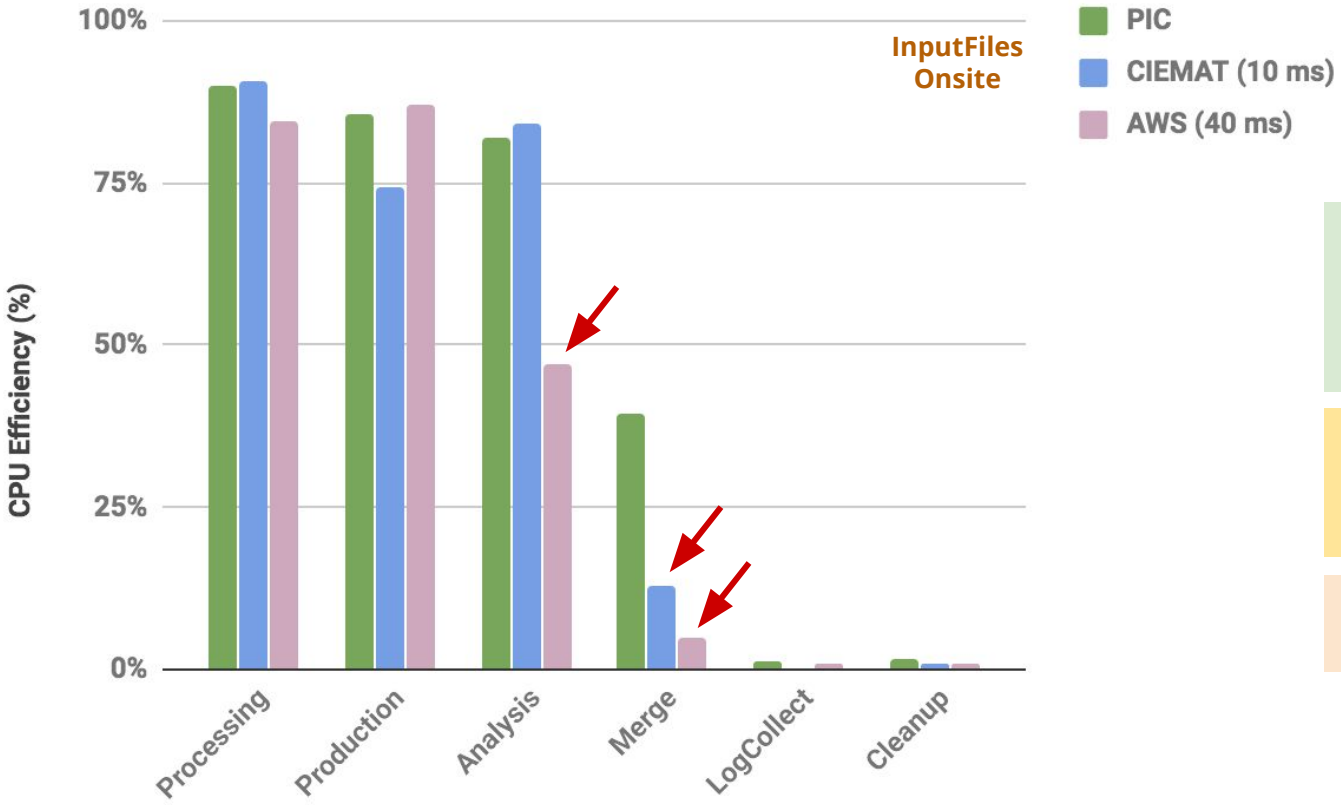
5.6 % of PIC jobs executed in CIEMAT

InputFiles Onsite



2.8 % of CIEMAT jobs executed in PIC

Towards a regional federation [CMS]



This tells us that within the region, PIC could run jobs either at PIC or CIEMAT (reading files from PIC), except merge (which should run locally always)!

This of course would cause an increase of PIC exports (stressing both for network and storage system… how much?)

At higher latencies (40 ms), analysis starts to be degraded

Data Access studies [CMS]

How are the **storage systems utilized** in PIC Tier-1 and CIEMAT Tier-2 for CMS? Are we working in the most optimal point?

~3% of data blocks are replicated both at PIC_Disk and CIEMAT, not an issue

Which data is susceptible to be **cached** and what could be the **benefits**? (we can simulate based on real data accesses)

PIC and CIEMAT are close enough (10ms) - shall we aim for a **data federation** or **consolidation** of storage in the region?

PhD student [C. Pérez Dengra] looking into this:

- In depth data access and performance studies, for both PIC and CIEMAT

→ Check contributions to pre-GDB - XCache (July 2019): talk#1 and talk#2

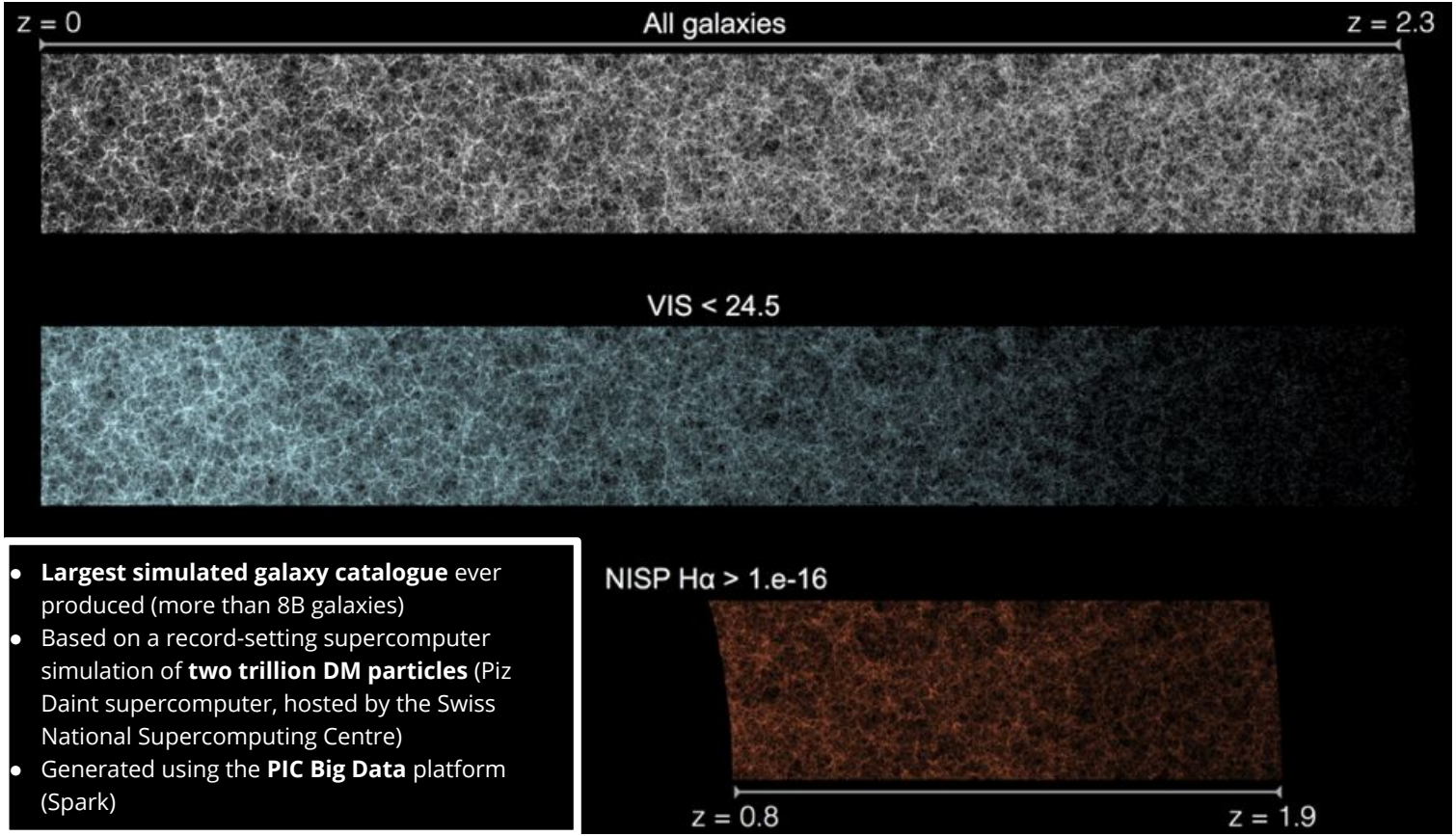
Data Access studies [CMS]

Including **CIEMAT Tier-2** and **CERN Tier-0** (collab. with CERN-IT) to draw conclusions at all Tier levels

These studies <u>can be done easily at any site running dCache</u> (since it gets data from the billingDB)

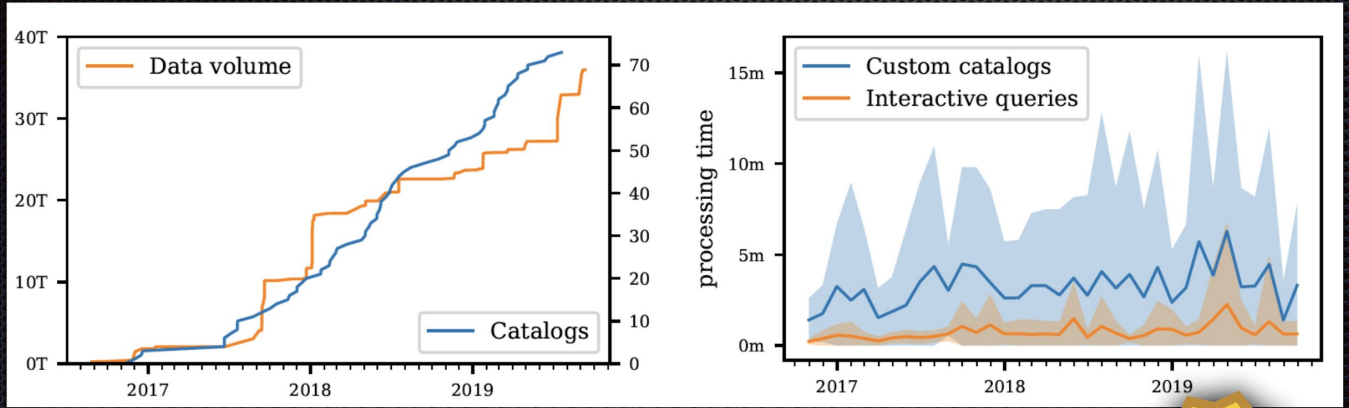
Contact us if interested!

Euclid Flagship mock galaxy catalogue



- **Largest simulated galaxy catalogue** ever produced (more than 8B galaxies)
- Based on a record-setting supercomputer simulation of **two trillion DM particles** (Piz Daint supercomputer, hosted by the Swiss National Supercomputing Centre)
- Generated using the **PIC Big Data** platform (Spark)

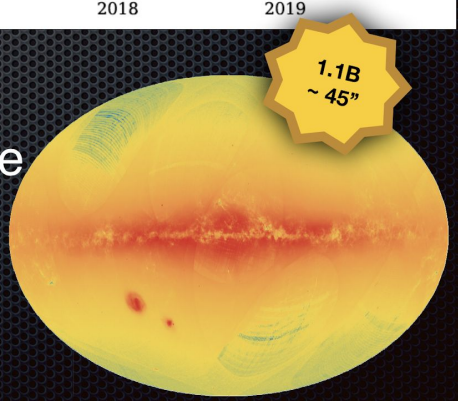
COSMO HUB on Big Data



- Web portal to perform interactive exploration and distribution of massive cosmological data based on Hive

https://cosmohub.pic.es

(Tallada et al. in prep)



Thanks!
Questions?

E. Acción, V. Acin, C. Acosta, Martin Børstad, J. Carretero, J. Casals, R. Cruz, M. Delfino, J. Delgado, J. Flix, F. López, G. Merino, C. Neissner, A. Pacheco, C. Pérez, A. Pérez-Calero, E. Planas, M.C. Porto, B. Rodríguez, P. Tallada, F. Torradeflot, A. Vedaee

www.pic.es