A Spanish data cache service for the CMS experiment

[Accepted project @ RES Datos 2021]

C. Acosta, F. J. Calonge, A. Delgado, J. Flix, J. M. Hernández,
C. Pérez, A. Pérez-Calero, A. Sikora

I Workshop de Computing y Software de la Red Española de LHC / Virtual
28-29 April 2021

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN
Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas
cfp CIEMAT física de partículas
PIC port d'informació científica

RES Data Services

Red Española de Supercomputación (RES) got the mandate from the Ministry to evolve to a Network of Supercomputing and Data.

There was an open call in spring 2019 for new Data nodes: **PIC/CIEMAT applied**

In September 2019, the RES council approved the new Data Network with 9 nodes

BSC	Barcelona Supercomputing Center - Centro Nacional de Supercomputación	Coordinación SC+D
COMPUTAEX	Fundación Computación y Tecnologías Avanzadas de Extremadura	SC+D
SCAYLE	Supercomputación Castilla y León	SC+D
CESGA	Centro de Supercomputación de Galicia	SC+D
IAC	Instituto de Astrofísica de Canarias	SC+D
CSUC	Consorci de Serveis Universitaris de Catalunya	SC+D
BIFI	Biocomputación y Física de Sistemas Complejos -Universidad de Zaragoza	SC+D
SCBI	Centro de Supercomputación y Bioinovación - Universidad de Málaga	SC+D
PIC	Port d'Informació Científica - CIEMAT	D

RES Data Services

PIC
port d'informació científica

1st call for data projects closed on Jan 28th 2021 - projects were approved by 13th April 2021

Recursos de almacenamiento ofrecidos

	Centro	Tipo de almacenamiento ofrecido	TB	Al 3r año
1	BSC	HSM, ficheros, cinta + disco	4000	23000
2	BIFI	CEPH, ficheros y objetos, disco	200	600
3	COMPUTAEX	GPFS y Lustre, objetos, disco y SSD	220	660
4	CESGA	Disco y cinta	200	400
6	CSUC	S3, ficheros y objetos, disco	220	660
5	IAC	VFS de Lustre, Disco	900	900
7	PIC	S3, Swift-CEPH, ficheros y objetos, disco y cinta	1200	2000
8	SCAYLE	S3, objetos, disco y SSD	400	1000
9	SCBI	S3, objetos	2000	2000
		Total PB	9.3	34.8

18 projects approved, for the period 2021-2023, with yearly allocations of storage resources

https://www.res.es/es/acceso-a-la-res/resolucion-convocatorias

[5 projects at PIC node]

RES Data Services

Lider	Título	Centro	Almacenamiento	Servicios adicionales
Pablo Loza-Álvarez	SLN intelligent datasets management	PIC	- 2021: 64 TB en disco y 64 TB en cinta - 2022: 122 TB en disco y 122 TB en cinta - 2023: 180 TB en disco y 180 TB en cinta - fin de proyecto: 354 TB en disco y 354 TB en cinta	
Javier Rico	Deploying the MAGIC Data Legacy	PIC	- 2021: 200 TB en disco y 200 TB en cinta - 2022: 200 TB en disco y 200 TB en cinta - 2023: 200 TB en disco y 200 TB en cinta	
José Flix Molina	A Spanish data cache service for the CMS experiment	PIC	- 2021: 100 TB en disco - 2022: 200 TB en disco - 2023: 200 TB en disco	Xcache
Christian Neissner	PAU survey: High-precision photometric redshift from narrow band observations in the visual range	PIC	- 2021: 75 TB en disco y 25 TB en cinta - 2022: 100 TB en disco y 100 TB en cinta - 2023: 150 TB en disco y 150 TB en cinta	50k horas de CPU por año
Andres Pacheco Pages	Experimental data storage of the ATLAS experiment from publications of the IFAE group	PIC	- 2021: 200 TB en disco y 200 TB en cinta - 2022: 200 TB en disco y 200 TB en cinta - 2023: 200 TB en disco y 200 TB en cinta	100k horas de CPU por año y servicios WLCG

https://www.res.es/es/acceso-a-la-res/resolucion-convocatorias

[5 projects at PIC node]

RES Data Services

Lider	Título	Centro	Almacenamiento	Servicios adicionales
Pablo Loza-Álvarez	SLN intelligent datasets management	PIC	- 2021: 64 TB en disco y 64 TB en cinta - 2022: 122 TB en disco y 122 TB en cinta - 2023: 180 TB en disco y 180 TB en cinta - fin de proyecto: 354 TB en disco y 354 TB en cinta	
Javier Rico	Deploying the MAGIC Data Legacy	PIC	- 2021: 200 TB en disco y 200 TB en cinta - 2022: 200 TB en disco y 200 TB en cinta - 2023: 200 TB en disco y 200 TB en cinta	
José Flix Molina	A Spanish data cache service for the CMS experiment	PIC	- 2021: 100 TB en disco - 2022: 200 TB en disco - 2023: 200 TB en disco	Xcache
Christian Neissner	PAU survey: High-precision photometric redshift from narrow band observations in the visual range	PIC	- 2021: 75 TB en disco y 25 TB en cinta - 2022: 100 TB en disco y 100 TB en cinta - 2023: 150 TB en disco y 150 TB en cinta	50k horas de CPU por año
Andres Pacheco Pages	Experimental data storage of the ATLAS experiment from publications of the IFAE group	PIC	- 2021: 200 TB en disco y 200 TB en cinta - 2022: 200 TB en disco y 200 TB en cinta - 2023: 200 TB en disco y 200 TB en cinta	100k horas de CPU por año y servicios WLCG

https://www.res.es/es/acceso-a-la-res/resolucion-convocatorias

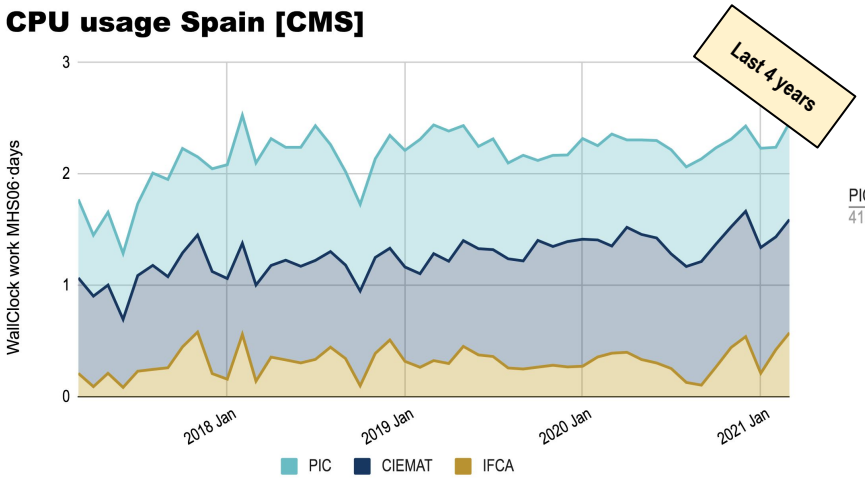
[5 projects at PIC node]

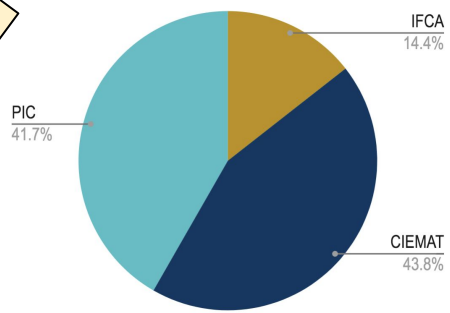
A Spanish data cache for CMS

CIEMAT contributes to WLCG with 2 data centers:

- one of the 13 **Tier1** centers around the globe (**PIC**), operated together with the IFAE
- one of the more than 100 WLCG **Tier2** centers, installed in Madrid at the **CIEMAT** site

Current CMS resources in Spain comprises around 5000 processing cores, 6 PB of disk space, and 9 PB of tape archival storage in PIC, CIEMAT and IFCA (Tier2)





A Spanish data cache for CMS

PIC
port d'informació
científica

One of the key elements of the evolution of the WLCG computing infrastructure towards HL-LHC will be the deployment of a **content delivery network** (CDN) for the direct streaming of data from the distributed storage system to the globally-spread processing compute resources

The proposed CDN can enable efficient remote data streaming directly into the processing application, temporarily caching relatively small amounts of input data close to the processing resource, and reducing data access latency by using read-ahead techniques

The technology for implementing the WLCG CDN has been developed during the last few years, based in the **XRootD framework**, which provides services and protocols to implement a distributed data storage federation, a caching layer (**XCache**) and an efficient network data access protocol

A Spanish data cache for CMS

The goal of this project is the deployment of the **Spanish WLCG CDN node for CMS**, setting up a **data cache service at PIC**, to serve data to processing and analysis jobs executed in the PIC Tier1 and the Spanish CMS Tier2 sites

- The data cache proposed in this project will achieve significant resource savings, both in **CPU** (improving processing efficiency) and **storage** (largely reducing data duplication needs). Infrastructure cost reduction is mandatory to cope with the much larger data volumes expected from the LHC over the next decade

The new remote data streaming model will require a sufficiently large wide area network (WAN) bandwidth. WLCG utilizes a private network, provided in Spain by **RedIris**, whose interconnections are planned to be upgraded in 2021 from the current **10 Gbps to (2)100 Gbps**, perfectly matching the change of architecture and the deployment timing

C. Pérez Dengra doing a PhD at PIC “Managing massive data of CMS Experiment: the evolution towards HL-LHC” ← topic aligned with his research activities

XCache: test instances (OSG repo)

In CIEMAT (XRootD 4.11):

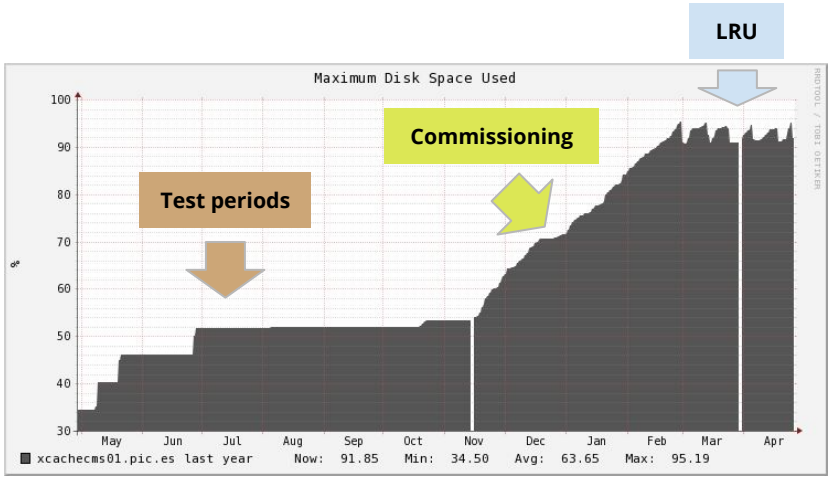
→ 22 TB, with RAID10, 8 cores, 8 GB RAM

→ Used by 2 Worker Nodes for all data tiers but only for TFC fall-back (data is not at CIEMAT)

In PIC (XRootD 5.1.1-1.3):

→ Distributed 4TBx36 [130 TB], no-RAID, 16 cores L5630 (HT enabled), 48 GB RAM, 10 Gbps - 90-95% occupancy

→ attached to one unique Worker Node that “caches all data tiers” as of today



150 TB of data copied in the period

The XCache applies LRU algorithm to flush the cache (95% → 90%)

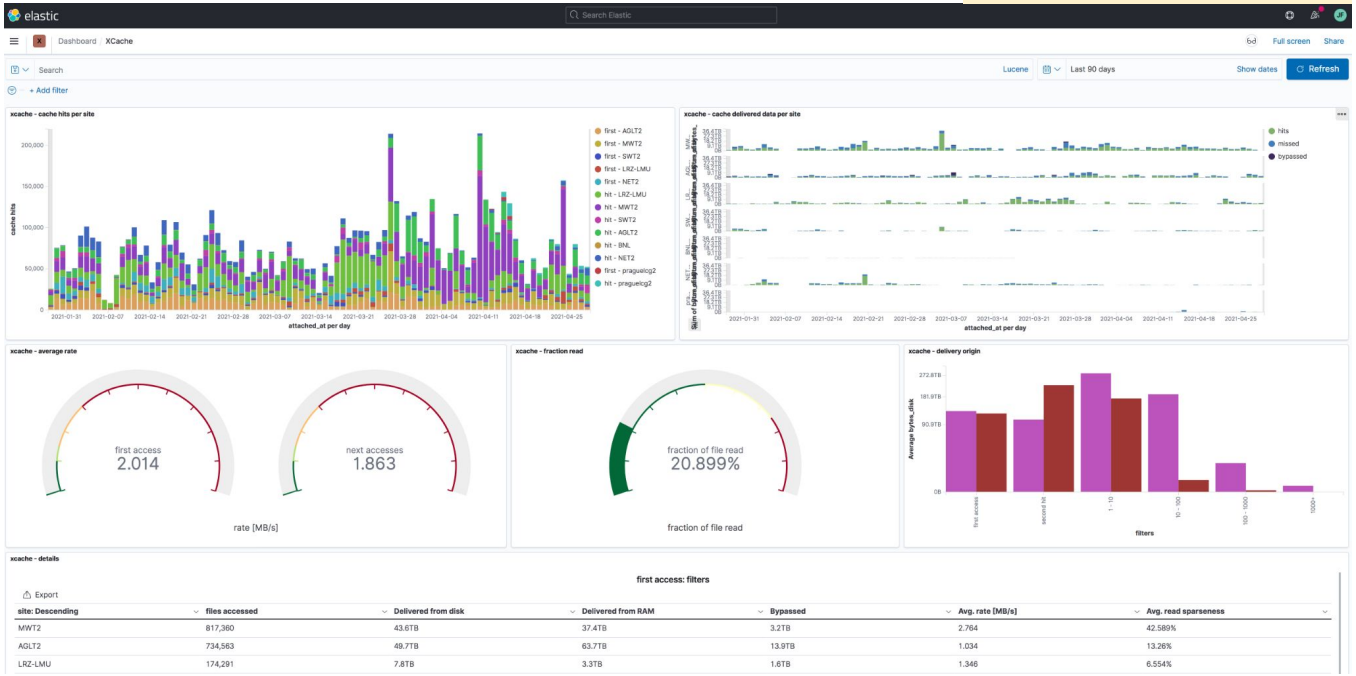
XCache: current status

Monitors and **proper settings** being deployed and/or investigated

- Local monitor (based on XRootD cinfo files)
- Information from CERN MONIT instances

e.g. ATLAS USA XCache dashboard

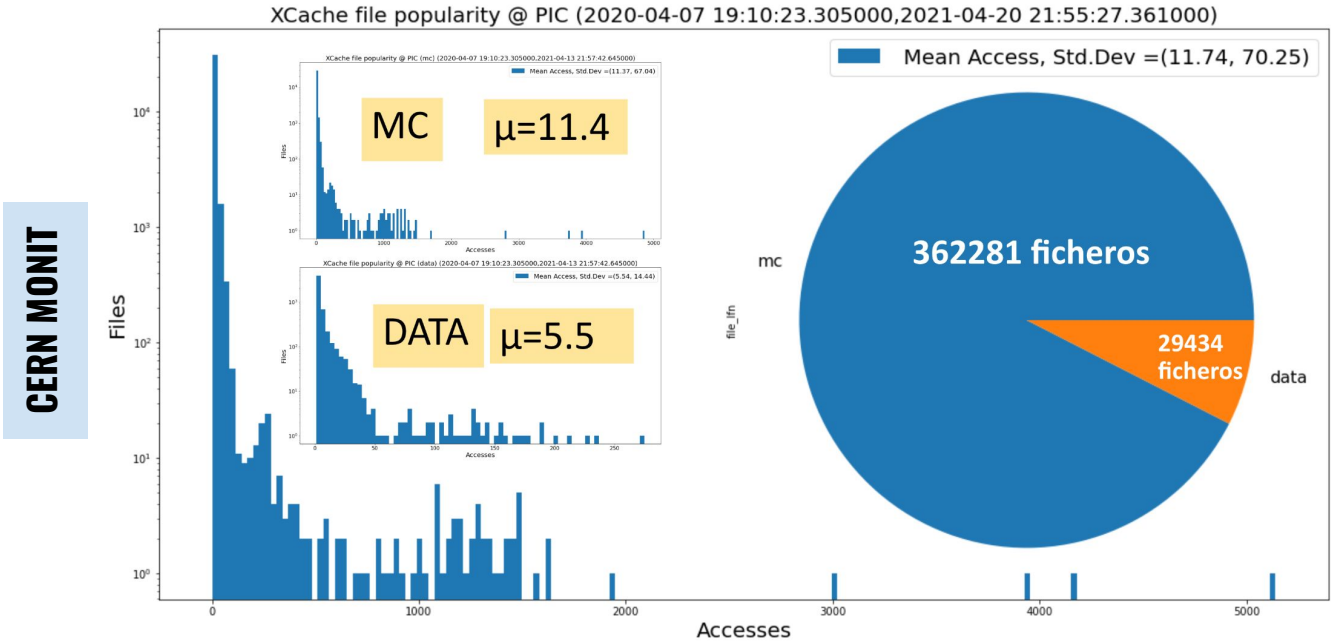
cinfo files



XCache: current status

Monitors and **proper settings** being deployed and/or investigated

- Local monitor (based on XRootD cinfo files)
- Information from CERN MONIT instances

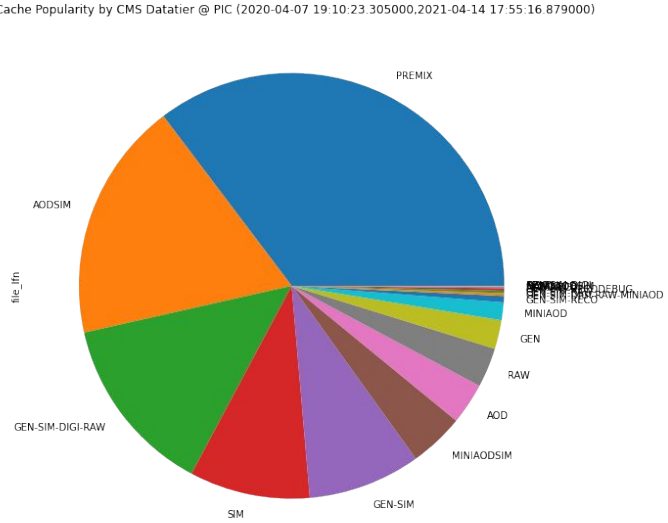


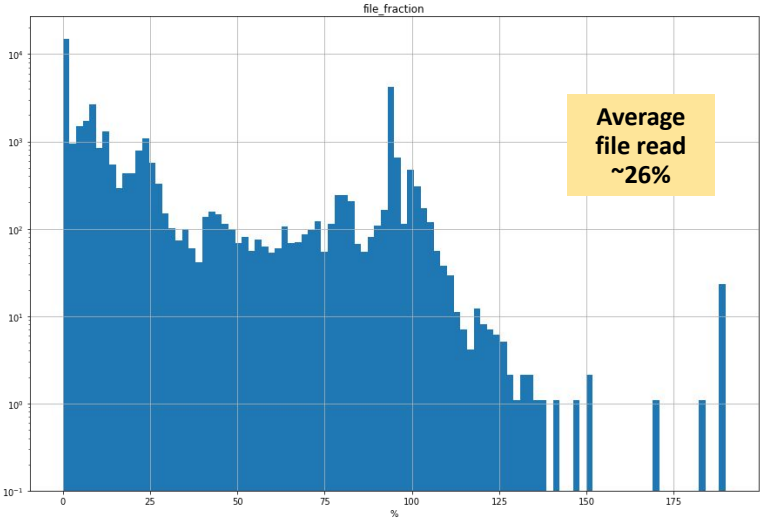
XCache: current status

Monitors and **proper settings** being deployed and/or investigated

- Local monitor (based on XRootD cinfo files)
- Information from CERN MONIT instances

CERN MONIT





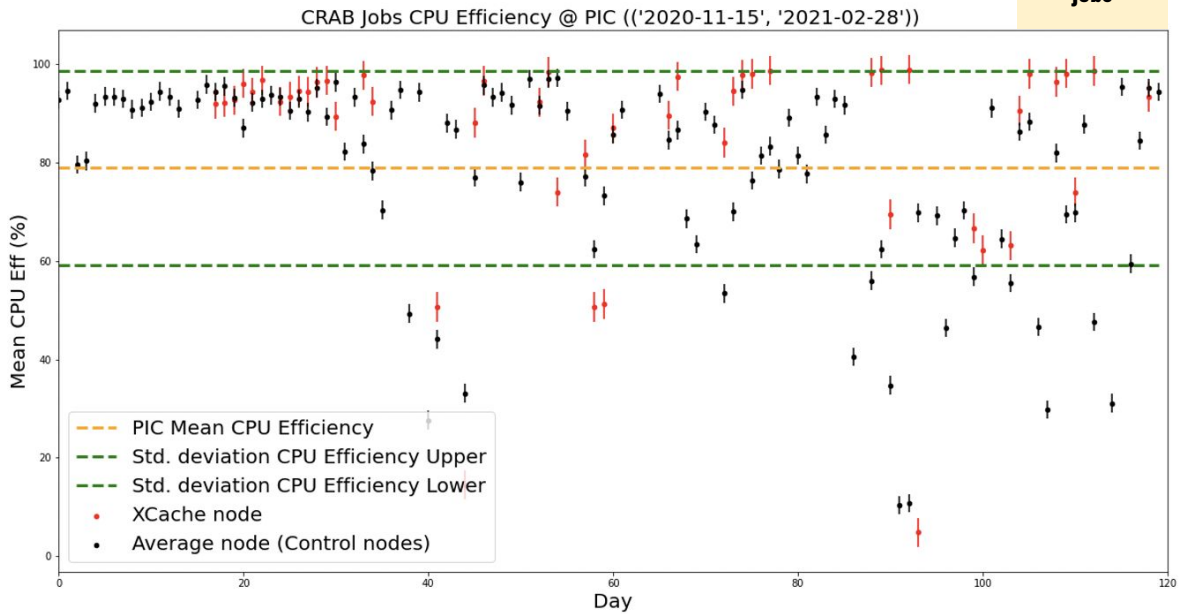
XCache: current status

Monitors and **proper settings** being deployed and/or investigated

- Local monitor (based on XRootD cinfo files)
- Information from CERN MONIT instances

CMS CRAB jobs

CERN MONIT



Xcache
μ_{CPUeff}=0.85

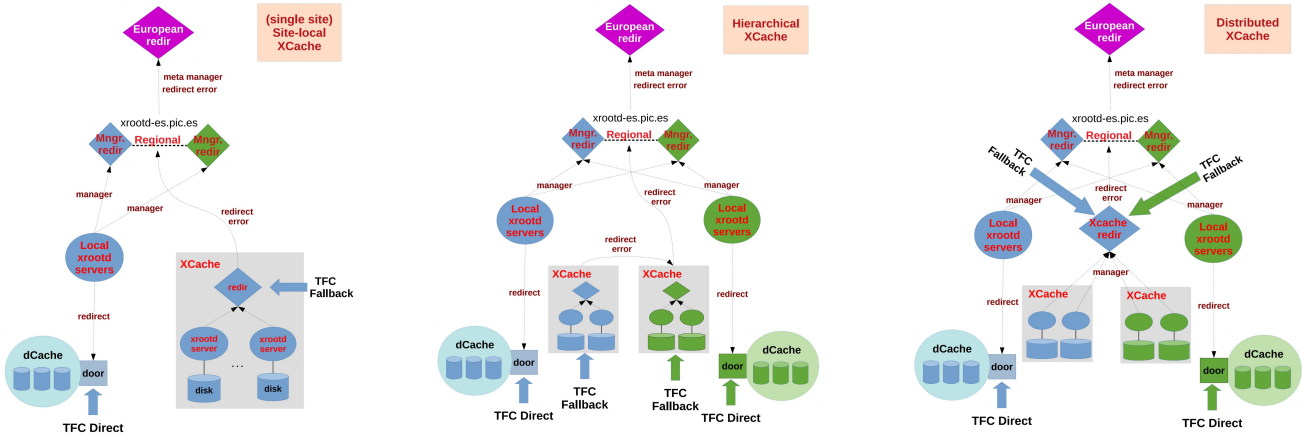
Average
μ_{CPUeff}=0.79

XCache: next steps

Evaluate the multi-site/regional XCache benefits in the region (interaction between XCaches?)

E.g: only **populate the Regional XCache** if the remote file to be read is off-region

- Allow local/regional reads w/o caching, profiting from the regional XRootD redirector deployed



Configure the XCache and the CMS TFC to get **files that are popular** (like AOD* files)

Enabling local monitoring (as any other production service at PIC)

Add **more Worker Nodes** in front of the XCache (eventually all) + **CIEMAT Worker Nodes**

Study the **improvement effects on the CPU efficiency** for those tasks using the XCache

Thanks!