

A content delivery network for CMS experiment in Spain

Carlos Pérez Dengra

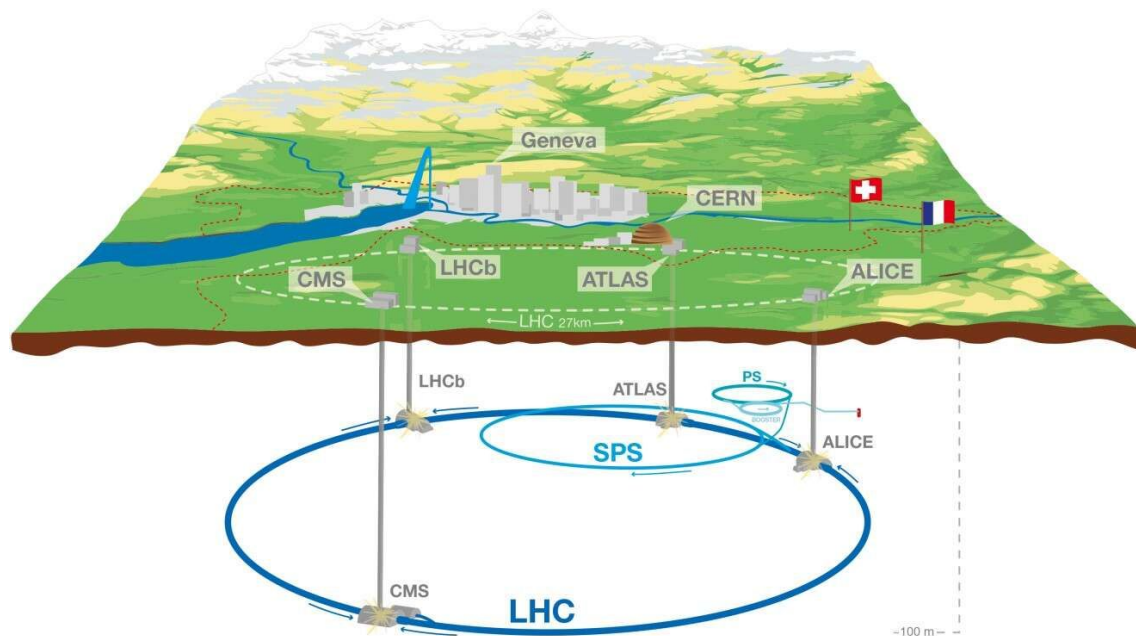
Port d'Informació Científica (PIC) - CIEMAT

JURES 2022, Cáceres. 14th-15th September of 2022

The Large Hadron Collider (LHC)

In the LHC, at CERN, two proton beams in form of bunches travel at the record centre-of-mass energy of 13.6 TeV in opposite directions.

These proton-proton collisions produce new particles, observed in four big detectors: **ALICE**, **ATLAS**, **CMS** and **LHCb**.

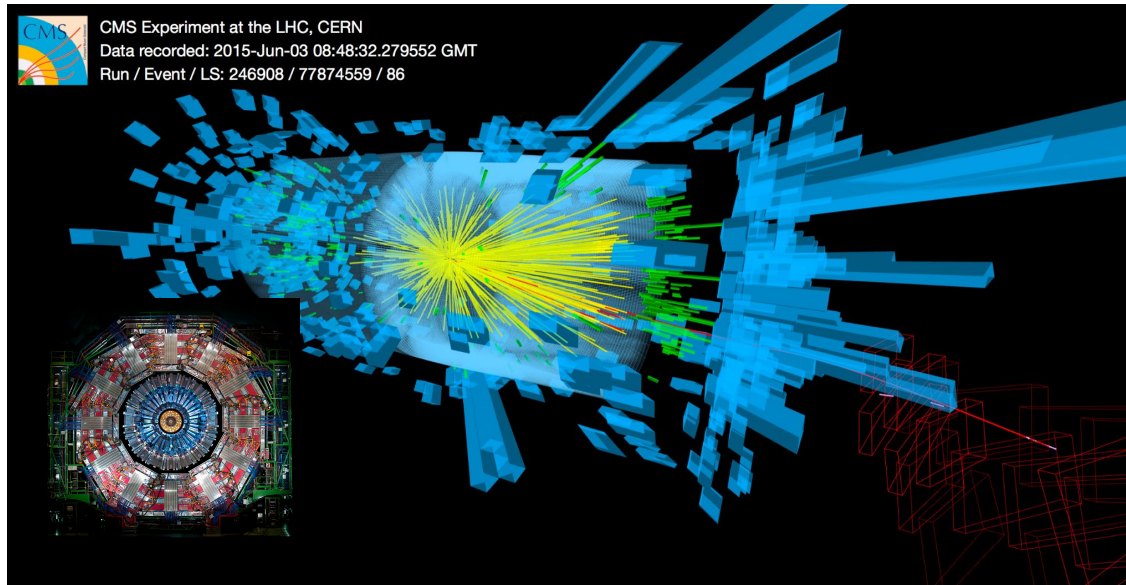


In 2012 ATLAS and CMS announced the discovery of the **Higgs Boson** that describes the mechanism in which particles can acquire mass

Data produced at the LHC

Proton collisions occur at 40M Hz at the center each of the LHC detectors, and data is collected at ~2k Hz (HLT trigger rates) at each experiment, yielding to up 60-80 PB of raw data in a nominal LHC year.

Acquired and simulated data is consequently stored at disk and tape systems within the LHC grid infrastructure.



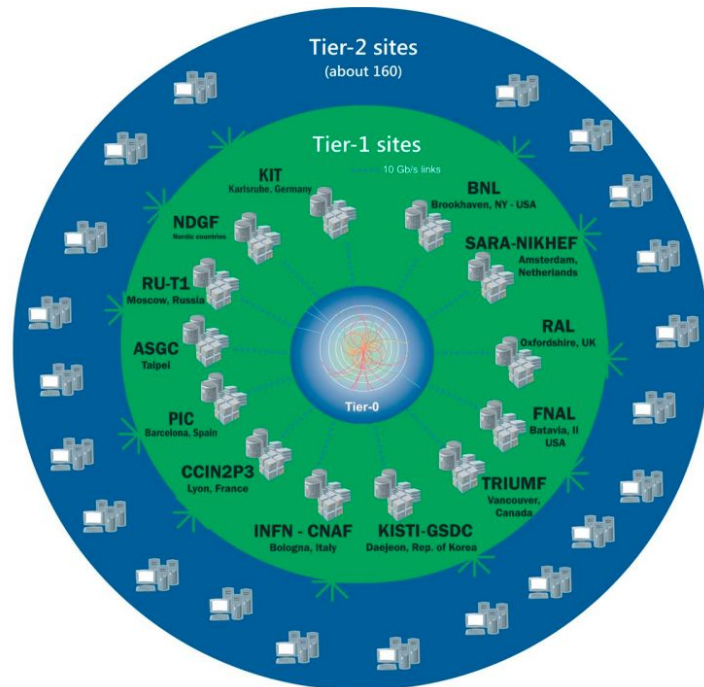
Worldwide LHC Computing Grid (WLCG)

LHC computing resources are distributed in ~170 sites in ~35 countries (the so-called WLCG).

Resources are usually offered in a hierarchical structure:

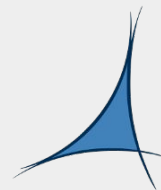
- **Tier-0 (CERN):** experimental data first processed / first copy of the raw data stored.
- **Tier-1s:** subsequent massive data processing , simulation campaigns / long-term preservation of data.
- **Tier-2s:** smaller facilities attached to Universities and research centers / end-users analysis data / simulation production.

LHC has currently **reached the Exabyte scale** among all experiments, and entered an active R&D phase towards the HL-LHC era (~2029)



PIC WLCG Tier-1

- PIC, created in 2003, is the WLCG Spanish Tier-1 site providing ~5% of Tier-1 data processing for CERN's LHC detectors ATLAS, CMS and LHCb.
- PIC is within the Spanish Supercomputing Network (RES) as part of RES-DATA.
- As a reference data center also supports neutrinos, astrophysics and cosmology projects.

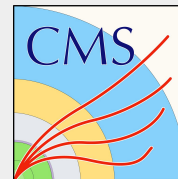


PIC
port d'informació
científica

CPU=10k CPU cores+GPUs
Disk=10 PB
Tape=35 PB

Tier-1
ATLAS, CMS, LHCb

**PIC Tier-1 Resources for CMS
experiment**



CPU ~22M hours/year
Disk=3.4 PB
Tape=11 PB

WLCG Computing challenges towards High-Luminosity LHC (HL-LHC)

WLCG computing infrastructure needs to evolve to face the new LHC era, the High-Luminosity LHC (HL-LHC) in 2029.

Several projects ongoing in Spain lead by PIC team and defined as strategic by RES :

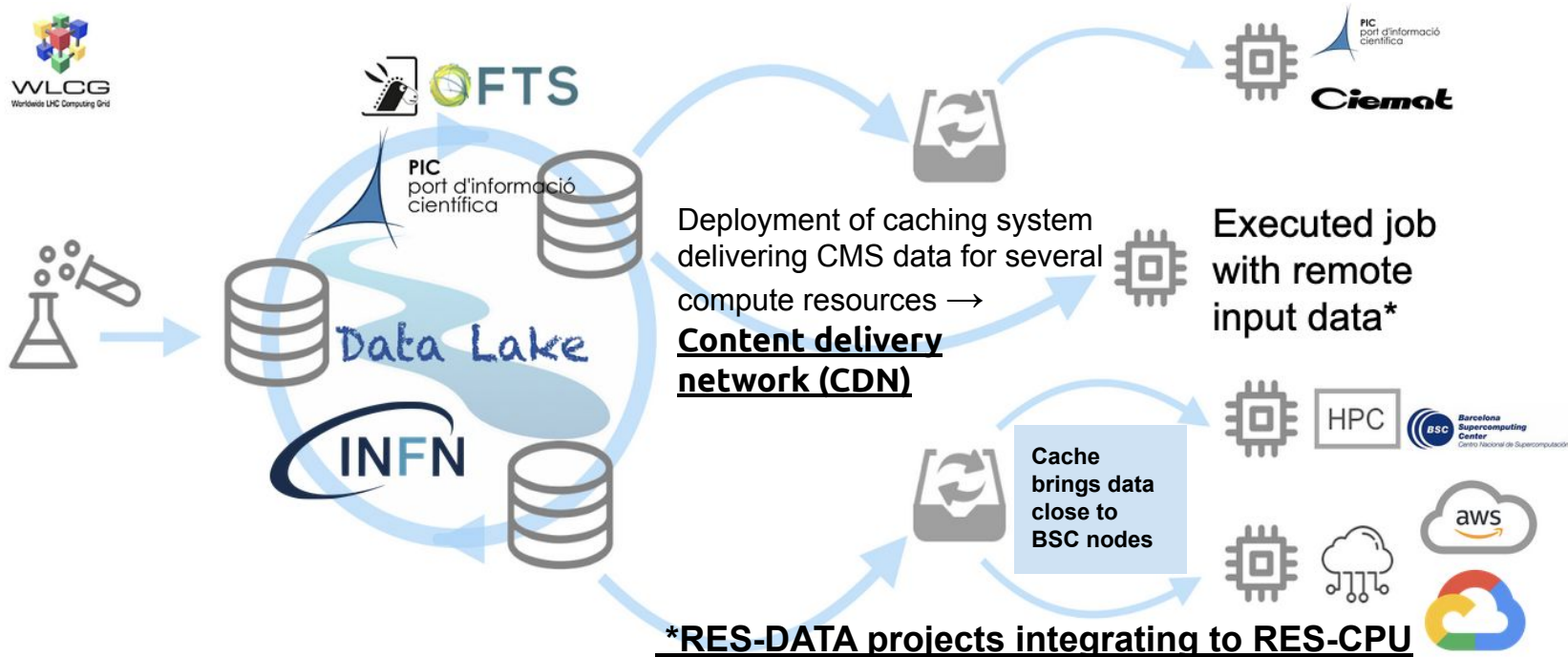
-RES-CPU: Integration of BSC resources to expand on CPU capabilities for LHC computing

-RES-DATA: Data Management evolution → cache evaluation for the CMS experiment to reduce storage costs and improve CMS jobs performances

The evolution of infrastructure will allow the sites to specialize in CPU or storage resources
→ based in **WLCG Data-lake model**.

This work is appreciated by the community and we are active contributors to these activities.

WLCG Computing challenges towards High-Luminosity LHC (HL-LHC)



Caching system deployed in CMS spanish Tier-1 and Tier-2

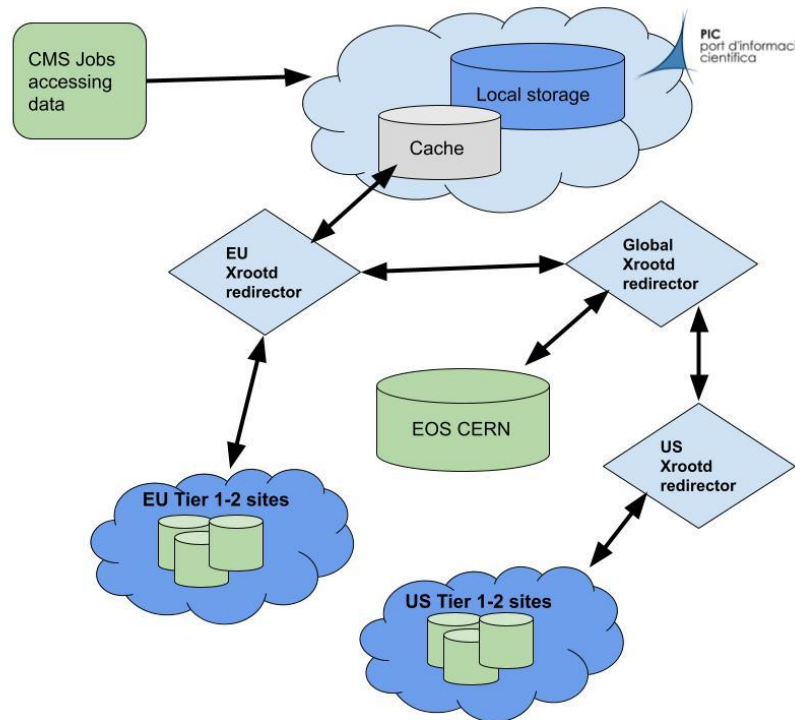
- PIC hosts a cache system for CMS data for the Spanish region. This is part of the project → ***“A Spanish data cache service for the CMS Experiment” (RES DATA-2020-1-0039).***

- Cache system is deployed through XCache service (xrootd protocol cache).

- Cache follows a Least Recently Used (LRU) deletion algorithm, deleting the least accessed files when it reaches a certain occupancy.

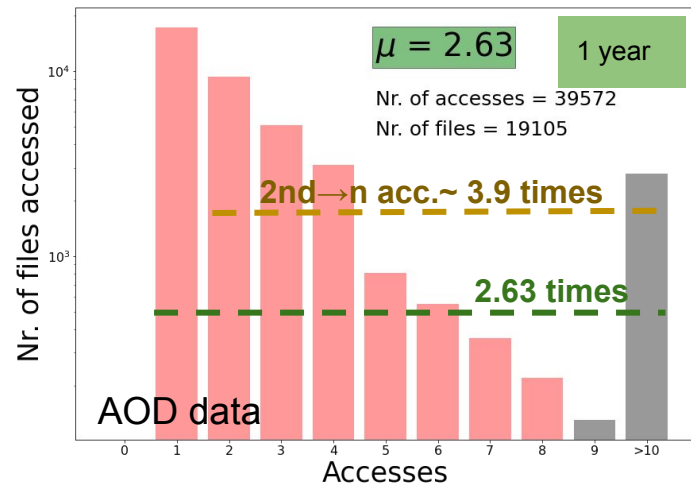
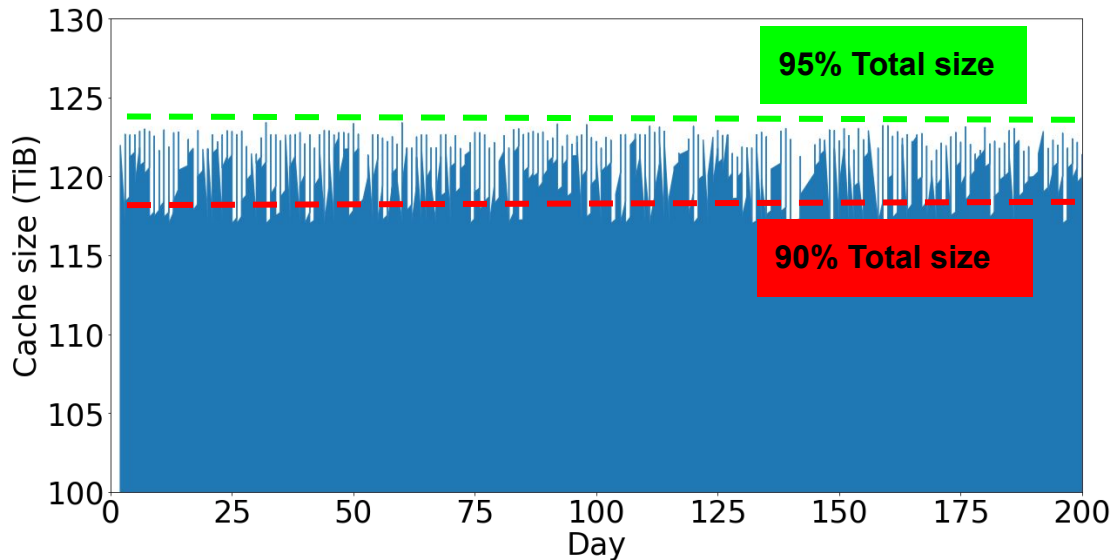
- CMS software allow tasks to read data remotely via xrootd protocol using redirectors → caching data would bring jobs data closer!

- **Reduce data access latencies**
- **Improve the global CPU efficiency**
- **Reduce bandwidth**
- **Reduce storage costs**



Current data status in caching system

- Total size deployed 150 TB, caching all remote data reads (not only analysis).
- LRU algorithm keep cache occupancy saturated between 95-90% of the total occupancy (PIC cache contains ~80k files).
- Average analysis file size is 2-4 GB → ~1k files monthly created and average re-accesses are ~4 times/file.



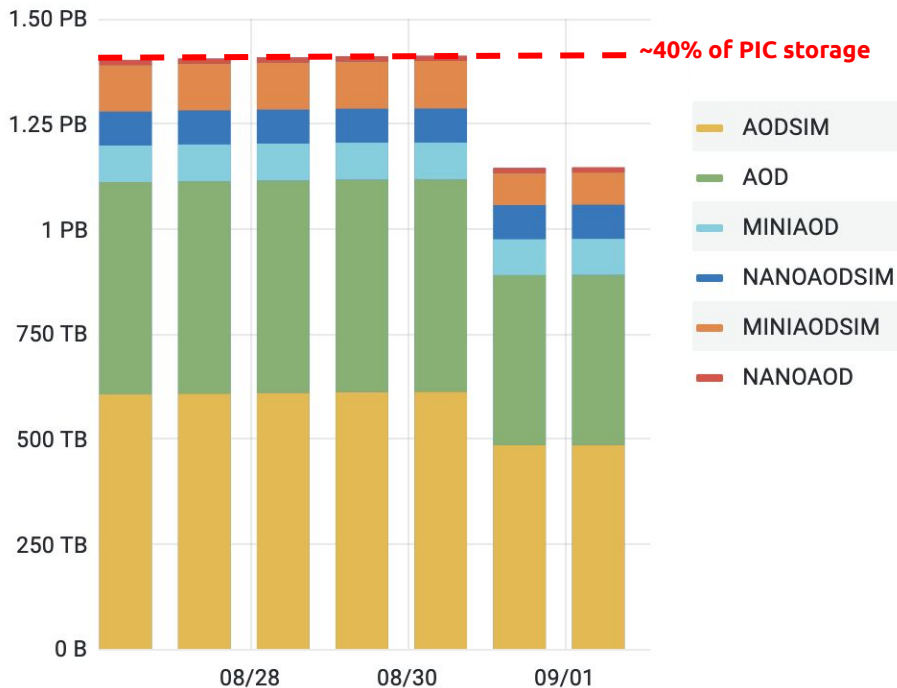
CMS jobs and storage usage at PIC

PIC executes about 100k CMS jobs per month (~2 MHours/month CPU)

~35% of CMS jobs correspond to analysis tasks that access reduced Analysis Object Data (AOD) formats, datasets that are typically re-accessed often.

Custodial and replica AOD samples co-exist in PIC storage. Currently, ~10% of CMS analysis jobs access remote input data.

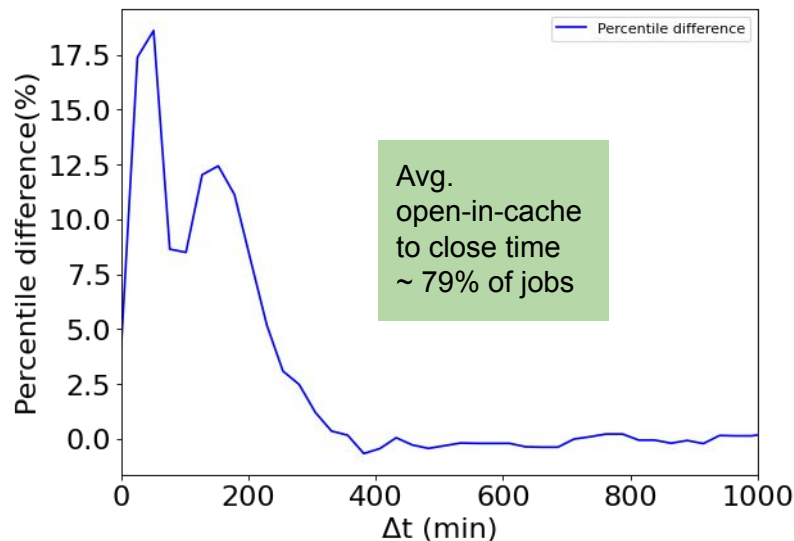
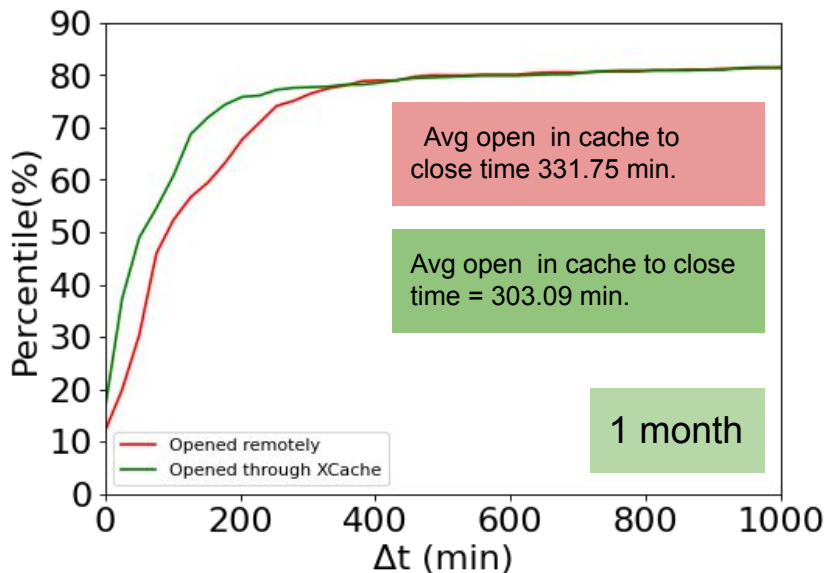
Deploying a well-sized cache would alleviate the storage costs, reducing the replicas that are present at the storage, and allowing efficient access to remote data.



Daily snapshots of the storage state are computed to evaluate the occupancy of disk by each type of data in the CMS Sites.

Impact on CMS jobs and storage

- Incorporating a cache into our infrastructure reduces the time it takes to access remote input data if it is already cached at PIC.
- The re-reads of files in the cache contribute to eliminate latency, improving the CPUeff.



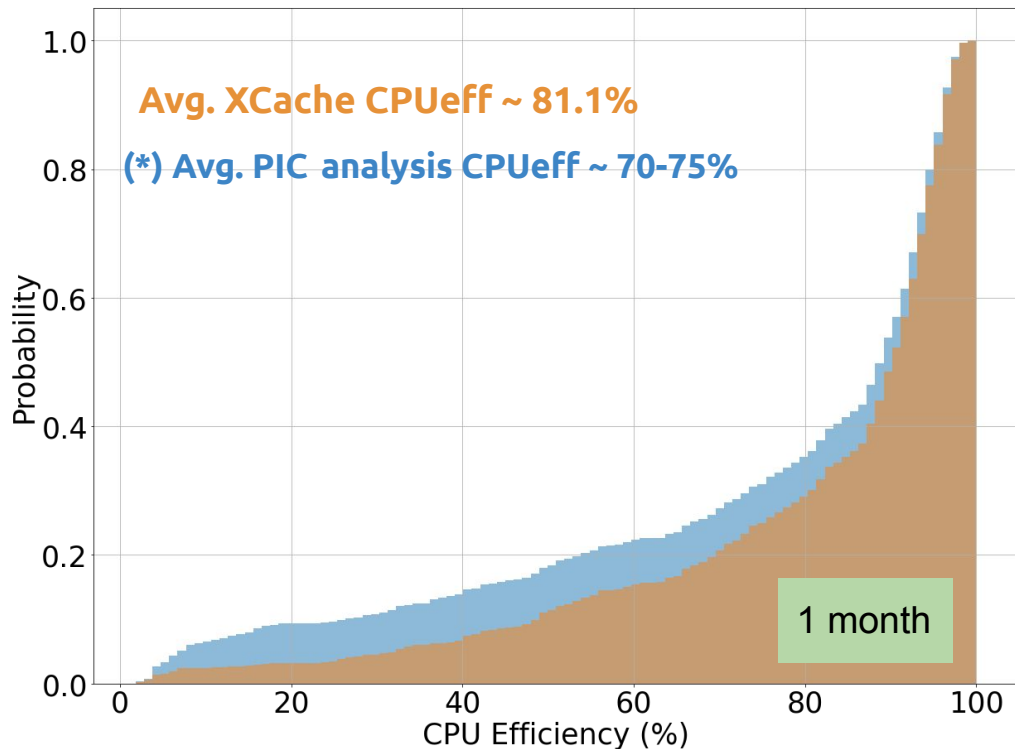
Accessing data in cache save up about 28.6 min/job of WallTime in average.

Impact on CMS jobs and storage

The 10% of time saved by accessing the data through the cache and not remotely, results in an improvement between 5-10% in the efficiency of the jobs.

The WallTime savings not only affect the costs of CPU usage time, but also the costs of network usage.

These results are computed using real jobs in production. To evaluate the adjusted CPU Eff gains, controlled jobs tests must be performed.



900 jobs analysed within a month

Outlook

- Cache system has been deployed at PIC with aim of storing part of the replicated CMS data kept in the PIC storage. CMS CIEMAT Tier-2 is placed in Madrid at low latency (9ms RTT). Due to the good interconnection at 100 Gbps, the aim is that the cache deployed at PIC would serve data to this site as well
→ CMS Spanish data federation.
- Also, caching different data types would increase the traffic of data to the cache, and would affect its dimension and performance. The system is still under evaluation, to better optimize and find the optimal working point


Conclusions

- PIC is integrating to its LHC computing infrastructure new elements: HPC resources and data management evolution, among others.
- 55Mhours of BSC resources have been spent by PIC's LHC experiments at 2021. 50% of the Spanish CPU pledge is used for simulation. It is the entire simulation that is expected to be executed in the BSC
- PIC as a CMS Tier-1 has incorporated a caching system based in xrootd protocol XCache as part of the data management evolution towards HL-LHC. This new element in the infrastructure will allow CMS community to save storage costs and improve the CPU efficiency by reducing the latency of accessing remote input data.
- The efficiency of jobs accessing analysis data through the cache improves by these means between 5-10% and saving 28.8 min/job of RTT, opening the door to expand these results to more types of experimental data.

Thank you for your attention!






Backup
slides



Use of BSC resources from LHC experiments in Spain

In 2020 an agreement was signed with BSC and LHC computing turned into one of the selected strategic projects at the HPC facility. Extensive LHC simulations required for ATLAS, CMS and LHCb are allocated, each quarter, in the BSC compute resources.

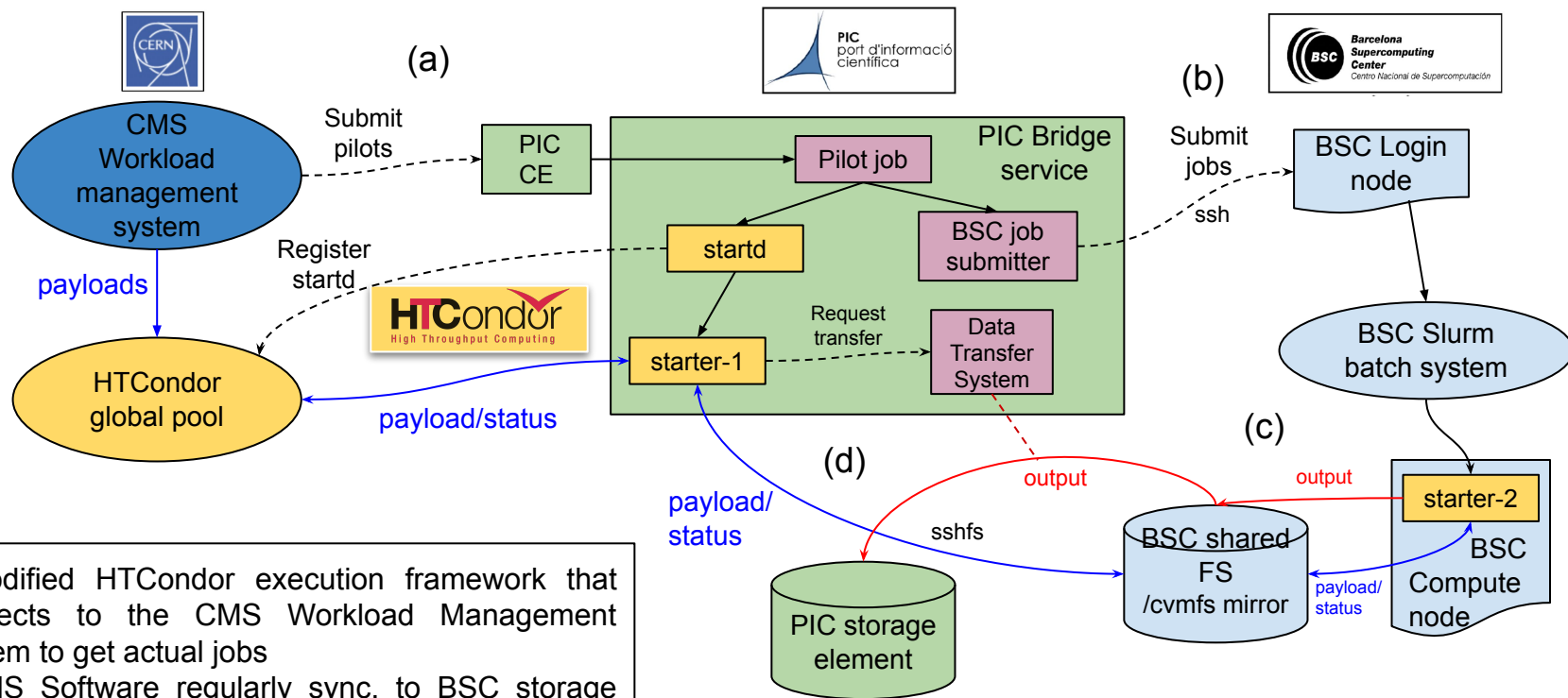
LHC Experiment			
Exploited in 2021 (CPU Mhrs)	27.2	18.2	9.4

Total ~55Mhrs for LHC simulation in 2021



Dedicated setups at PIC were deployed, for each experiment, using SSH protocol for both job submissions and transferring data between PIC and BSC.

Use of BSC resources for CMS experiment at PIC



- Modified HTCondor execution framework that connects to the CMS Workload Management System to get actual jobs
- CMS Software regularly sync. to BSC storage system
- Data Management system integrated with the executed jobs to transfer outputs to PIC