21st International Workshop on Advanced Computing and Analysis Techniques in Physics Research

Deploying a cache content delivery network for CMS experiment in Spain

Carlos Pérez Dengra (PIC-CIEMAT)[1], **José Flix Molina** (PIC-CIEMAT), **Anna Sikora** (Autonomous University of Barcelona)
on behalf of the CMS Collaboration

Abstract: *The CMS experiment at the Large Hadron Collider (LHC) uses a hierarchy of redirectors based on XRootD protocol that allows execution tasks to run by accessing input data that is stored on any Worldwide LHC Computing Grid (WLCG) site. In 2029 the LHC will start the High-Luminosity LHC (HL-LHC) program, when the luminosity will increase in a factor 10 as compared to the current values. This scenario will also imply an unprecedented increase of simulation and collision data to transfer, process and store in disk and tape systems. Two Spanish WLCG sites that support CMS, the PIC CMS Tier-1 and the CIEMAT CMS Tier-2, have have been exploring content delivery network type solutions in the Spanish region. One of the possible solutions under development has been the deployment of caches between the two sites that store the data requested by the jobs remotely, so that they get closer to the nodes to improve their job efficiency and input data transfer latency. In this contribution, we analyze the impact of deploying physical caches in production between PIC and CIEMAT, as well as the impact they have on job efficiency.*

XCache for CMS in PIC Tier-1 and CIEMAT Tier-2

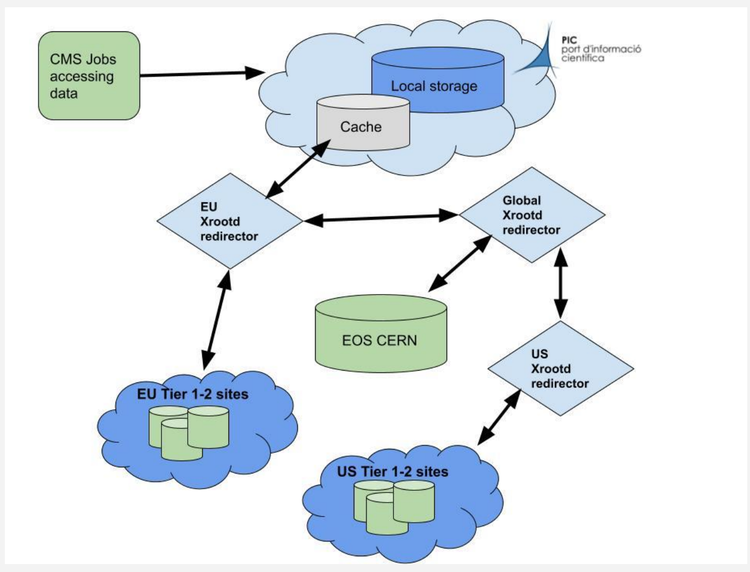

Figure 1: XCache configuration for remote input data reads of CMS jobs in PIC Tier-1.

PIC Tier-1 and CIEMAT Tier-2 are hosting a cache system for CMS data based on XRootD protocol: XCache. CMS Any Data, Anytime, Anywhere (AAA) XrootD protocol allows redirecting access to data required by jobs to the closest site where it is available, within a hierarchy of redirectors, if they are not stored in the local site. The configuration of XCache for PIC Tier-1 (as well as for CIEMAT Tier-2) is shown in the **Figure 1**. When a job is executed in the compute nodes and input data is not found in the local disk storage, the cache downloads it accessing the data on-demand, instead of reading it from any other WLCG Tier site. This configuration is being explored, as part of the envisioned R&D tasks towards HL-LHC computing challenges**[1]**.

Cache status: a brief snapshot

PIC CMS Tier-1 has currently deployed a 150TB dedicated disk server with XCache service. The same service, with 22TB, is deployed at CIEMAT CMS Tier-2 as well. Both are configured with Least recently used (LRU) algorithm to keep saturated the occupancy between 95% and 90% of the total volume at both sites. Caches store CMS collision and simulated data. **Table 1** shows the current status of the distribution of cached data for both sites.

	PIC CMS Tier-1 Cache size = 150 TB		CIEMAT CMS Tier-2 Cache size = 22 TB	
	Collision data	Monte-Carlo data	Collision data	Monte-Carlo data
Volume (TB)	44.74	74.06	7.16	13,79
% Of total	62	38	66	34

Table 1: Cache data type occupancies by October of 2022 at PIC CMS Tier-1 and CIEMAT CMS Tier-2.

Controlled submission of analysis jobs accessing MiniAOD files

Analysis files are the most accessed data by CMS users elsewhere. Hence, **MiniAOD** samples are the most suitable files to be placed in caches**[2][3]**. Tasks analyzing a sample MiniAOD file have been submitted in a controlled production-like environment, and reading data from PIC cache and remote sites, in order to understand the effects on the CPU efficiency gains.

Events processing through XCache

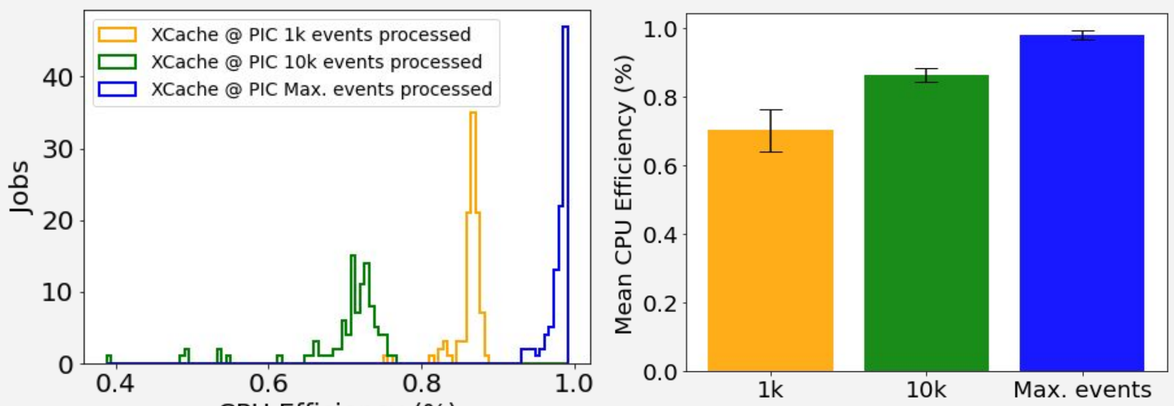

Figure 2: CPU efficiency considering different number of events processed by analysis jobs accessing data from XCache @ PIC.

We studied what is the effect on the CPU efficiency with the number of events processed by the test job in the selected MiniAOD file (110k total events). As seen in Figure 2, the maximum CPU efficiency is reached when the total number of events are processed, as expected given the overheads of CMS Software (CMSSW) initialization phase. Hence, we decided to use the total number of events (110k) in our studies.

CPU efficiency accessing data from PIC and CIEMAT

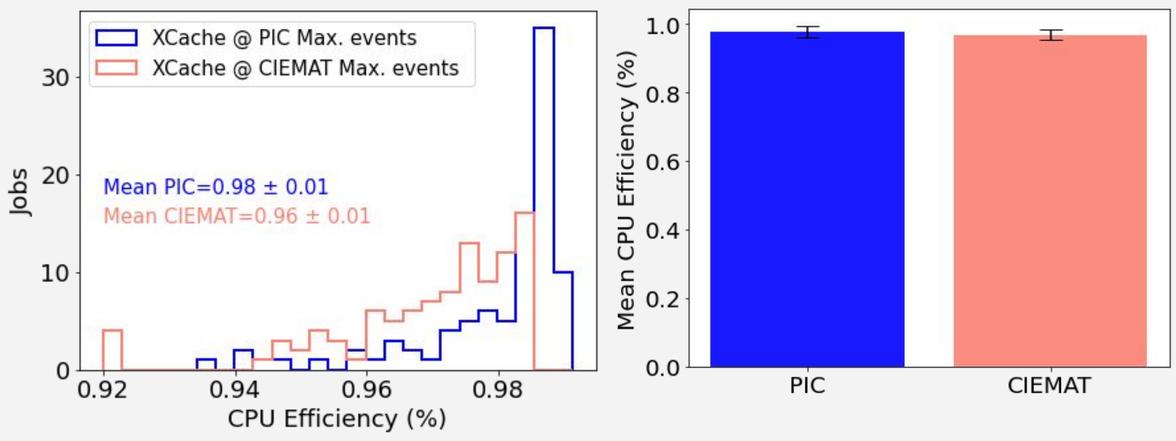

Figure 3: CPU efficiency for the test jobs executed in PIC and accessing input data from XCache @ PIC and XCache @ CIEMAT.

A series of test jobs have been executed at PIC reading the selected MiniAOD data from the PIC and CIEMAT caches, respectively. The relative degradation on the CPU efficiency is not very significant, of about 2%, when running the tasks reading data from CIEMAT cache, as compared to the results obtained when reading data from the PIC local cache, as seen in **Figure 3**. The two sites are separated by ~500 km and with a measured latency of ~9ms. Previous studies that re-routed production jobs between the two sites also showed no significant degradation on the CPU efficiency for not I/O intensive tasks**[4]**.

CPU efficiency of jobs accessing data from farther sites

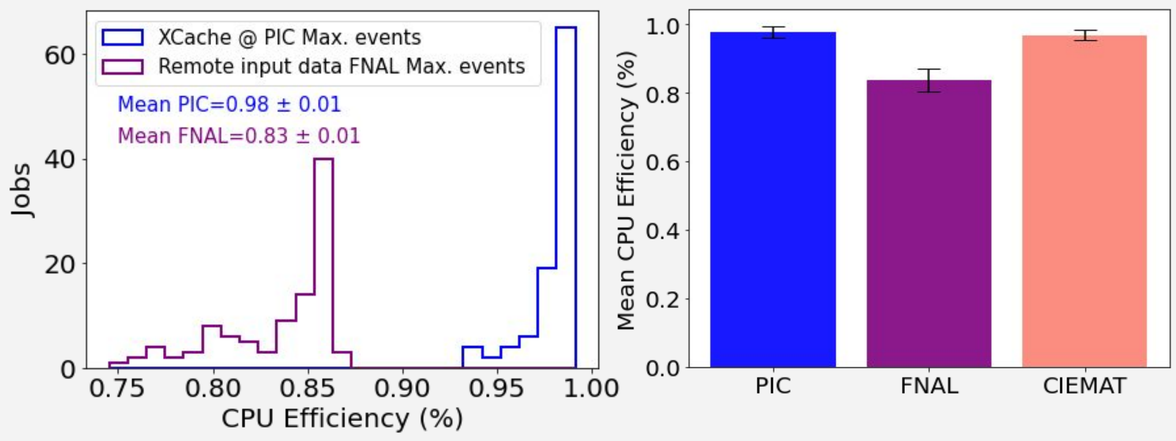

Figure 4: Figure 4: CPU efficiency distribution for the test jobs executed in PIC accessing input data from XCache @ PIC, FNAL CMS Tier-1 and CIEMAT CMS Tier-2.

The same test was run in PIC environment, but this time reading the selected MiniAOD data from the FNAL Tier-1 site, which is located in Chicago (USA). The latency of ~150 ms between PIC and FNAL is expected to degrade the mean CPU efficiency of the jobs accessing remote input data. **Figure 4** shows the observed relative degradation on the CPU efficiency of about 15%, as compared to the results obtained when reading data from the PIC local cache.

Conclusions & Outlook

Studies about possible gains in CPU efficiency for CMS tasks using caching systems in PIC Tier-1 and CIEMAT Tier-2 sites have been carried out for different conditions. A controlled execution of analysis jobs has been chosen, accessing the same MiniAOD file placed at the local PIC cache and at other remote sites.

We observed no significant degradation on the CPU efficiency when data is served either from a local cache or a cache within the Spanish region. Hence, a single cache placed in PIC Tier-1 could serve data to all of the Spanish CMS Tier-2 sites. When reading files from remote sites overseas (FNAL), we do observe significant degradations on the CPU efficiency. In this case, accessing the popular data files through a cache presents an efficient solution.

Further tasks will include higher statistics, with more sites included, and using other AOD files to evaluate the degradation of CPU efficiency of these job types as a function of latency.

References

[1] CMS Offline Software and Computing. *CMS Phase-2 Computing Model: Update Document.* Tech. rep. Geneva: CERN, Jul. 2022. URL: https://cds.cern.ch/record/2815292
[2] Delgado Peris, A., Flix Molina, J., Hernández J., Pérez-Calero Yzquierdo, A., Pérez Dengra, C., Planas, E., Rodríguez Calonge, J., Sikora, A 2019 "CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2", EPJ Web Conf., 245 04028 (2020) DOI: 10.1051/epjconf/202024504028.
[3] Pérez Dengra, C., Flix Molina, J., Sikora, A 2022 "New storage and data access solution for CMS experiment in Spain towards HL-LHC era", 20th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT), Virtual and IBS Science Culture Center, Daejeon, South Korea, 29-3 December 2021, viewed 14th October of 2022.
[4] C. Acosta-Silva, A. Delgado Peris, J. Flix, J. M. Guerrero, J. M. Hernández, A. Pérez-Calero Yzquierdo, F. J. Rodriguez Calonge, J. Gómez del Pulgar Ruano A 2019 "Lightweight site federation for CMS support", EPJ Web Conf. 245 03013 (2020) DOI: 10.1051/epjconf/202024503013.

* The authors of this work thanks the PIC and CIEMAT teams for their support in these studies and by deploying novel cache services for the CMS experiment in the Spanish region.

**This project is partially financed by the Ministry of Science and Innovation of Spain MINECO, within the grants for projects with references FPA2016-80994-C2-1-R, PID2019-110942RB-C22, DATA-2020-1-0039 and BES-2017-082665. It has also been supported by Ministerio de Ciencia e Innovación MCIN AEI/10.13039/501100011033 under contract PID2020-113614RB-C21 and by the Generalitat de Catalunya GenCat-DIUiE (GRC) project 2017-SGR-313. Finally, this project is also supported by the Red Española de Supercomputación (RES) within the grant DATA-2020-1-0039.

[1]Contact: cperez@pic.es