Deploying a cache content delivery network for the CMS experiment in Spain

Carlos Pérez Dengra (PIC-CIEMAT)[1],
José Flix Molina (PIC-CIEMAT)[2],
Anna Sikora (Autonomous University of Barcelona)[3]
on behalf of the CMS Collaboration.
[1,2] Carrer de l'Albareda Edifici D · Campus UAB · 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain. [3] Carrer de les Sitges Edifici Q · Campus UAB · 08193 Bellaterra (Cerdanyola del Vallès), Barcelona, Spain.

[1]-cperez@pic.es, [2]-jflix@pic.es, [3]-anna.sikora@uab.cat

Abstract: The High-Luminosity Large Hadron Collider (HL-LHC) is set to begin in 2029, with an expected tenfold increase in luminosity. This increase will generate a significant amount of simulation and collision data that must be processed, stored, and transferred across the Worldwide LHC Computing Grid (WLCG) sites. Despite the LHC's almost 15-year operating experience, handling such a large amount of data requires a re-evaluation of data management techniques, and several R&D activities have emerged in response. The Spanish WLCG sites supporting the Compact Muon Solenoid (CMS) experiment, the PIC Tier-1 and the CIEMAT Tier-2, are exploring the use of data caches. These data caches are deployed to bring popular datasets closer to compute nodes where applications are executed. CMS software enables the execution tasks to remotely read data via XRootD redirectors. By deploying data caches, data access latency could be reduced, and CPU efficiency of tasks could be improved. To evaluate the feasibility of deploying an efficient content delivery network in the region, controlled testbeds have been used in Spain. A real CMS analysis job was executed to understand the benefits of caching data as compared to remote data reads. These studies aim to contribute to the ongoing R&D activities for future LHC data management. Overall, the deployment of data caches could significantly improve the management and processing of the data generated by the HL-LHC, leading to better scientific outcomes.

1. Introduction

The Large Hadron Collider (LHC) **[1]** started operations in 2009. The participating experiments at the facility have generated so far 1 exabyte of simulated and collision data from proton and ion collisions. These data are stored and processed using the Worldwide LHC Computing Grid (WLCG) **[2]**, a globally distributed computing infrastructure with 170 centers in 35 countries. The infrastructure is divided into Tier-0 center at CERN, 13 Tier-1 centers, and around 150 Tier-2 centers worldwide. As data production and computational demands are expected to increase significantly during the High-Luminosity LHC (HL-LHC) period, the LHC computing experts are actively exploring innovative solutions to address these unprecedented demands of compute resources, and introduce new techniques which would flatten these resources increases, in order to keep the resources under the budget constraints of the experiments. To minimize the cost of storage and improve data delivery at scale, the LHC experiments have launched an R&D program in collaboration with other current and future data-intensive experiments (mainly in HEP and Astroparticle Physics) facing similar computational challenges **[3]**.

To achieve efficient cost operations and reduce storage resources, a new scheme is being considered that would involve consolidating storage resources into fewer sites within the WLCG, reducing data duplication and profiting from cost-effective large deployments of storage resources. Additionally, content delivery network (CDN) solutions, such as caching systems, can be employed to further reduce the amount of storage required in different regions and to bring data close to CPU-only centers. The four major experiments at the LHC collaborate in the WLCG DOMA working group **[4]** to explore innovative ways to manage the LHC experiment data, and in particular the CMS experiment **[5]** data. Among all the proposed outcomes, data federations based in the XCache **[6]** service have been proved to be successful use cases **[7][8]**. XCache is the preferred caching system for XRootD **[9]**, the framework used for fast and efficient access to distributed data storage systems in most of the High-Energy Physics (HEP) experiments. This novel technology has been deployed for its adoption into the data management of the experiment after several efforts made by the community that are aligned with the goals outlined in the WLCG strategy plan towards HL-LHC. Data caches reduce data transfer over the wide-area network and decrease data access latency.

The objective of the study presented in this contribution is to evaluate the CPU efficiency gains in jobs executed in the Spanish CMS sites when using the XCache service. Popular datasets are those related to analysis activities, hence cache services have been deployed in the two sites, and dedicated tests with a real analysis job have been performed to compare the effects of reading data remotely or served from the local caches. Both sites are separated by ~620 km (~9 ms latency), hence we have as well studied the effects of deploying a single cache in PIC Tier-1 serving data to the whole region.

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

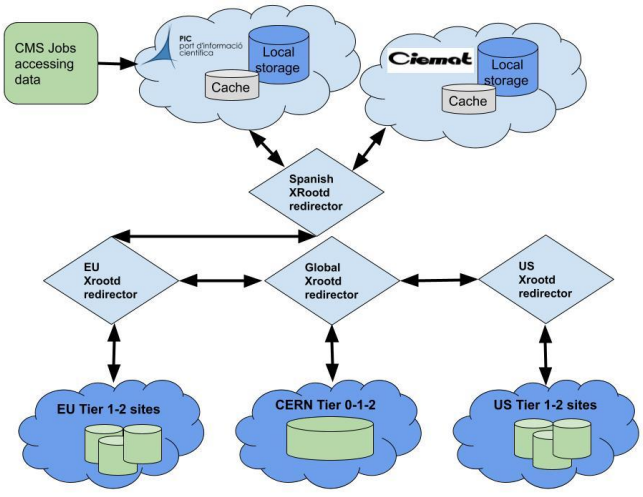

Figure 1: XRootd redirectors infrastructure with the inclusion of the XCache service in PIC and CIEMAT sites.

Figure 1 shows the XRootD redirectors infrastructure that spans all of the CMS Tier-1 and Tier-2 sites in Europe and the United States. The hierarchical subscription of XRootD redirector servers provides resiliency in data access towards clients. If the accessed data is not available at the site where the CMS job is executed, the client will be automatically redirected using the hierarchy of XRootD redirectors to the storage server where that data exists, and data will be served from a remote site at task execution time. Despite the suitability of this system, the use of the network and latency effects for distant sites can have negative effects on tasks' CPU efficiency. The XCache service allows a site or regional

network to cache data frequently used by the CMS experiment, reducing data transfer over the wide-area network and decreasing access latency, hence potentially minimizing the aforementioned effects. The XCache system operates by caching data on-demand when a job is executed and when the required data is not found in the local disk storage. The service can be as well configured to cache specific types of data, and let other data types use the XRootD redirector infrastructure.

Two XCache services have been deployed at the PIC and CIEMAT sites. Both sites opted for a single storage server, with 150 TB and 22 TB capacity, respectively. In PIC, the storage server disks are exposed individually, while at CIEMAT a RAID6 system has been adopted. In an initial stage, both XCaches were configured to cache all types of CMS data. Although this is not the optimal running mode, since we want to cache popular datasets only, this configuration proved to be useful to better tune the systems at scale. In particular, running at cache saturation allowed us to set the proper low and high usage "water marks" at the cache, set to 90% and 95%, respectively **[10]**. When usage goes above the high "water mark", the XCache service deletes cached files until usage goes below the low "water mark". These levels ensure that popular data is kept at the cache, maximizing the hit rates (data re-reads). A Least Recently Used (LRU) algorithm is used to identify files that are suitable for deletion, and deletions are performed when the high "water mark" is reached.

3. Controlled submission of analysis jobs accessing MiniAOD files

Among all types of data generated by the CMS experiment, the final analysis files (aka, AOD file types) are the most frequently accessed data by the users of the CMS computing infrastructure. To optimize the performance and efficiency of the XCache system, it is crucial to carefully consider the types of files that are most suitable for caching, and study the potential benefits when using the service. Based on previous research and usage patterns **[11][12]**, MiniAOD files accessed by user's analysis jobs have been identified as the most appropriate files to be stored in cache . The CMS AOD files have a reduced set of reconstructed physics objects for higher-level analysis, and MINIAOD files contain only the relevant information for faster processing and quick analysis, with the latter being the most accessed when performing final analyses for publications.

In order to evaluate the potential benefits of caching MiniAOD samples, a series of controlled jobs have been executed in a production-like environment in the region. For that purpose, an analysis template job **[13]** has been executed to access MiniAOD data, with or without the XCache infrastructure deployed at PIC and CIEMAT. This muon Physics Object Group (POG) job performs a tag and probe analysis for the events stored in a MiniAOD file. The tag/probe method involves a loop over all of the events in the file and then selecting a "tag" particle that satisfies specific selection criteria, which may include identifying it as a particular type of particle or requiring it to have a specific energy range. For these tests, we placed the input files at both caches deployed in CIEMAT and PIC, and we also accessed these files remotely from FNAL, which is a Tier-1 computing center located in Chicago (USA), that was also holding these input data in their storage system.

Executing tasks have calls outside the main event loop (initialization or writing the output file, for example). While smaller number of events result in shorter stage-out times and outputs, the overhead of the execution task does not significatively depend on the number of processes events. This overhead is well determined, since it occurs before and after the event loop. A preliminary study was conducted to set the maximum number of events to analyze from the selected MiniAOD file, since the main event loop dominates the CPU efficiency of the task. The analysis template job runs in single-core mode and it takes around 28 seconds to initialize, before entering the processing event loop. After multiple times execution, the task takes 6.1 ± 0.2 HS06·hour **[14]** to process the complete 110,323 events present in

the template MiniAOD file (with size of 2.9 GB), with a CPU efficiency of 98.35 ± 0.05 %. The maximum memory usage of the application is 1.47 ± 0.01 GB, and the average input file read throughput during job execution is 2.46 ± 0.02 MB/s. Figure 2 (left) shows how the CPU efficiency reaches a plateau for the number of events processed, and Figure 2 (right) shows the input file read throughput evolution with the events processed. For this study we decided to use the total events in the file. With these scores, we conclude that the selected analysis template job is not an I/O intensive task.

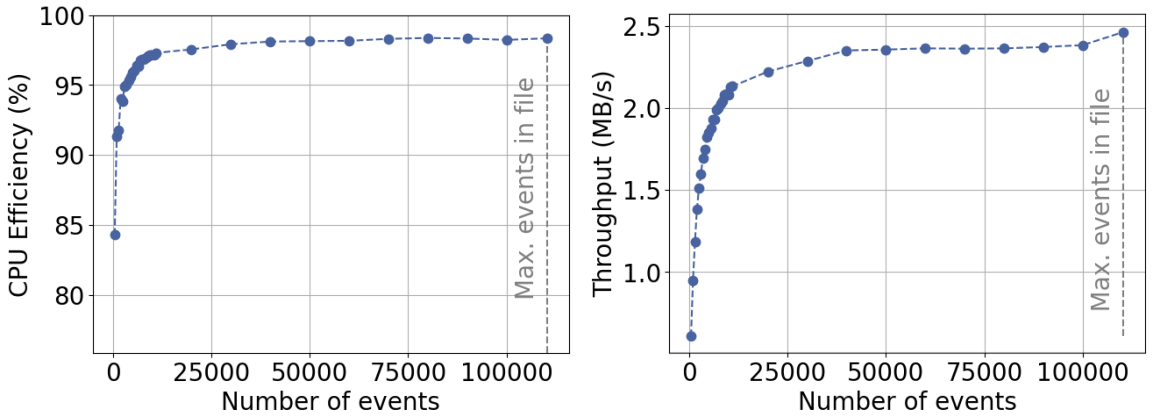

Figure 2: CPU efficiency (%) and events throughput (MB/s) as a function of the total number of events processed, from a task executed in PIC compute node reading from local XCache.

4. Effects on the CPU efficiency when using caches or performing remote reads

We studied the CPU efficiency for a series of analysis jobs (serially) executed at PIC, reading the input MiniAOD data files from both the PIC and CIEMAT caches. The serialization of the job executions ensured that only one job was accessing the input data at a time, and the tests were carried out in an empty compute node, in order to make the tests run isolated from other factors that could alter the results. Despite a separation of approximately 620 km between the two sites, and a measured latency of 9ms, the results showed a minor degradation in CPU efficiency of approximately 2% when reading data remotely from the CIEMAT cache (97.33 ± 0.14 %), as compared to reading locally from the PIC cache (98.35 ± 0.05 %) in Figure 3. Indeed, this observation aligns with previous studies that have demonstrated that the CPU efficiency of non I/O intensive tasks is not significantly impacted when jobs are rerouted between the two sites **[15]**.

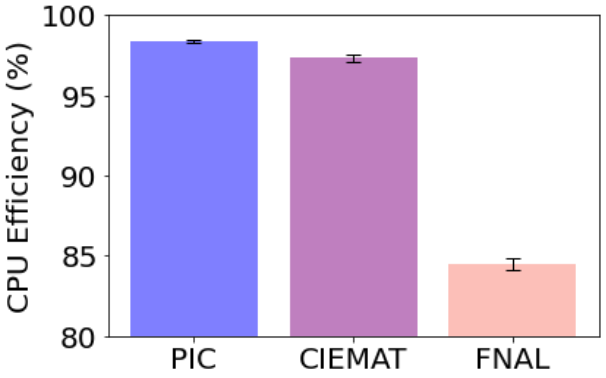

Figure 3: CPU efficiency distribution for the test jobs executed in PIC accessing input data from XCache at PIC Tier-1, XCache at CIEMAT CMS Tier-2 and FNAL CMS Tier-1.

A similar test was performed by executing the analysis test tasks in PIC compute node, in the same configuration as stated above, but this time reading the input MiniAOD files as stored at the FNAL Tier-1 site in Chicago, USA, which is approximately 7,077 km away from PIC center. The measured latency of 150 ms between PIC and FNAL is sufficient to result in a degradation of the mean CPU efficiency of these jobs. As also shown in Figure 3, the results revealed a significant degradation of approximately 14% in CPU efficiency (84.46 ± 0.34 %), as compared to reading the input data from the PIC local cache. While remote access to data across the transatlantic network is not a conventional CMS practice, we conducted the study to assess the advantages of bringing data from distant locations closer to compute nodes.

5. Conclusions and outlook

The results of our study provide valuable insights into the potential for improving CPU efficiency for CMS tasks using caching systems. We found that accessing MiniAOD data from local caches at PIC and CIEMAT, or a single cache within the Spanish region, did not result in a significant performance degradation. Our findings suggest that a single cache placed in PIC Tier-1 could effectively serve data to all Spanish CMS Tier-2 sites without a significant impact on performance. When reading data from remote overseas sites like FNAL, we observed a significant decline in CPU efficiency for these analysis tasks. This indicates that increased latency can result in a substantial degradation of CPU efficiency in the large-scale data and distributed systems we made use of. The XCache service can be used to catch and store popular data files, and would provide a more efficient solution for the executed tasks in the region. To note that ~25% of the PIC and CIEMAT compute power is used to compute analysis tasks, hence improving the CPU efficiency might have a positive outcome, a benefit that is subject to an evaluation. Using XCache would also reduce the amount of storage resources deployed in the region. The current AOD types CMS datasets in PIC and CIEMAT comprises 65% of the total storage usage (4.6 PB), and we plan to understand the scale of the XCache service in the region which would be needed to alleviate part of the storage that holds analysis files. We additionally plan to expand these studies to include a higher number of (remote) sites, the use of a wider range of AOD types, and the use of I/O intensive tasks, whose CPU efficiencies could be more affected by higher read access latencies. This will enable us to get a more comprehensive understanding of the impact of latency on the degradation of CPU efficiency for tasks accessing popular datasets. These ongoing efforts contribute to the understanding of the use and the benefits of the XCache service in the Spanish region, and ultimately within the CMS computing infrastructure.

References

[1] The Large Hadron Collider (LHC) Homepage: https://home.cern/science/accelerators/large-hadron-collider. Last accessed 4 May 2020.
[2] WLCG project, Last accessed on 6th June of 2016: http://wlcg.web.cern.ch/.
[3] J. Albrecht, et al, "A Roadmap for HEP Software and Computing RD for the 2020s", Computing and Software for Big Science volume 3, Article number: 7 (2019) DOI:10.1007/s41781-018-0018-8.
[4] X. Espinal, et al, ”The Quest to solve the HL-LHC data access puzzle. The first year of the DOMA ACCESS Working Group”, International Conference on Computing in High Energy and Nuclear Physics (CHEP), Adelaide, Australia, 4-8 november 2019, viewed 5 November 2019.
[5] S. Chatrchyan et al. “The CMS Experiment at the CERN LHC”. In: JINST 3 (2008), S08004. DOI:10.1088/1748-0221/3/08/S08004.
[6] L.A.T Bauerdick and K. Bloom and B. Bockelman and D.C. Bradley and S. Dasu and J.M. Dost and I. Sfiligoi and A. Tadel and M. Tadel and F. Wuerthwein and A. Yagil and the CMS collaboration “XRootd, disk-based, caching proxy for optimization of data access, data placement and data replication”,
Journal of Physics: Conference Series. 513-4 042044 (2014) DOI:10.1088/1742-6596/513/4/042044.
[7] D. Ciangottini et.al, “Integration of the Italian cache federation within the CMS computing model.” (2019) DOI:014. 10.22323/1.351.0014.

[8] Edgar Fajardo, Matevz Tadel, Justas Balcas, Alja Tadel, Frank Würthwein, Diego Davila, Jonathan Guiang, Igor Sfiligoi "Moving the California distributed CMS XCache from bare metal into containers using Kubernetes". EPJ Web Conf. 245 04042 (2020). DOI: 10.1051/epjconf/202024504042.
[9] XRootD Homepage, https://xrootd.slac.stanford.edu. Last accessed 31 Jan 2023.
[10] Pérez Dengra, C., Flix, J., Sikora, A., on behalf of the CMS Collaboration, "Simulating a content delivery network solution for the CMS experiment in the Spanish WLCG Tiers**",** International Symposium on Grids & Clouds (ISGC 2022) oral presentation, https://indico4.twgrid.org/event/20/contributions/1116/. Last accessed 21st Feb 2023.
[11] 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.
[12] Pérez Dengra, C., et.al, 2023 *J. Phys.: Conf. Ser.* 2438 012053. DOI 10.1088/1742-6596/2438/1/012053.
[13] Ramírez Sánchez, Gabriel. "MuonAnalysis-MuonAnalyzer", GitLab, version 1.0, 2023, https://gitlab.cern.ch/garamire/muonanalysis-muonanalyzer/-/tree/master/.
[14] HEPiX Benchmarking Working group, https://w3.hepix.org/benchmarking.html. Last accessed 31 Jan 2023.
[15] 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.

Acknowledgements
The authors of this work express their gratitude to the PIC and CIEMAT teams for their support in these studies and for deploying novel cache services for the CMS experiment in the Spanish region. This project is partially financed by the Spanish Ministry of Science and Innovation (MINECO) through grants FPA2016-80994-C2-1-R, PID2019-110942RB-C22 and BES-2017-082665, which include FEDER funds from the European Union. It has also been supported by the Ministerio de Ciencia e Innovación MCIN AEI/10.13039/501100011033 under contract PID2020-113614RB-C21, the Catalan government under contract 2021 SGR 00574, and the Red Española de Supercomputación (RES) through the grant DATA-2020-1-0039.