Optimal XCache service for the CMS experiment in Spain

José Flix[1,2], Carlos Pérez[1,2], Anna Sikora[3], Paula Serrano[3], for the CMS Collaboration

[1]CIEMAT, Basic Research, Scientific Computing Unit, 28040, Madrid, Spain
[2]PIC, 08193, Bellaterra (Barcelona), Spain
[3]Universitat Autònoma de Barcelona (UAB), 08193 Bellaterra (Barcelona) Spain

E-mail: jflix@pic.es

Abstract. The Large Hadron Collider at CERN in Geneva is poised for a transformative upgrade, preparing to enhance both its accelerator and particle detectors. This strategic initiative is driven by the tenfold increase in proton-proton collisions anticipated for the forthcoming high-luminosity phase scheduled to start by 2029. The vital role played by the underlying computational infrastructure, the World-Wide LHC Computing Grid, in processing the data generated during these collisions underlines the need for its expansion and adaptation to meet the demands of the new phase. The provision of these computational resources by the worldwide community remains essential, all within a constant budgetary framework. While technological advancements offer some relief for the expected increase, numerous research and development projects are underway. Their aim is to bring future resources to manageable levels and provide cost-effective solutions to effectively handle the expanding volume of generated data. In the quest for optimised data access and resource utilisation, the LHC community is actively investigating Content Delivery Network (CDN) techniques, aiming to enhance the performance of executing tasks by facilitating the efficient reading of input data via caching content near the end user. A comprehensive study is presented to assess the benefits of implementing data cache solutions for the Compact Muon Solenoid (CMS) experiment for the Spanish compute facilities, playing a crucial role in supporting CMS activities. Data access patterns and popularity studies suggest that user analysis tasks benefit the most from CDN techniques. Consequently, a data cache has been introduced in the region to acquire a deeper understanding of these effects. In this contribution we will focus on the remote data accesses from users that execute tasks in the Spanish CMS sites, in order to simulate and discern the most optimal requirements in both size and network connectivity for a data cache serving the whole Spanish region. This is a mandatory step towards a better understanding of the service to be deployed in a federated fashion in the region.

1 Introduction

In response to the increasing demands for computing power and storage during the LHC high luminosity (HL-LHC) period, the LHC experiments have initiated an extensive R&D program [1]. This program aims to optimize existing tools and develop innovative solutions for data management and processing capabilities to maximize scientific output from the vast data generated during the HL-LHC phase, expected to start by 2029. Projections of necessary computing resources for the HL-LHC, compared with estimated resource availability under a flat-budget model, suggest that an annual resource increase of 10-20% will be insufficient to meet HL-LHC demands. Without novel ideas, future computational requirements in the World-Wide LHC Computing Grid (WLCG) may not be met.

Many ongoing R&D efforts focus on enhancing computing power through GPU integration, partial application vectorization, and using opportunistic resources and HPC centers. Storage services face complex challenges, prompting a new approach: the WLCG Data-Lake model [2, 3]. This model aim to consolidate storage resources into fewer WLCG sites serving smaller centers through simplified data caches. Inspired by Content Delivery Networks (CDNs), this approach optimizes costs through strategically placed caches near high data demand points. High-bandwidth networks interconnecting WLCG sites, such as LHCONE and LHCOPN [4], must handle intense data flows within the Data-Lake. Ongoing R&D focuses on network virtualization and programmable networks [5] to ensure agile, cost-effective infrastructures. Lightweight storage systems, specifically data caches, will support both traditional (Grid) and opportunistic (Cloud/HPC) compute resources, ensuring flexibility and scalability. Efficient data caching mechanisms close to end users will enhance system responsiveness and user experience.

2 The CMS Context

The default protocol for CMS (Compact Muon Solenoid [6]) jobs involves processing data at its designated location. However, CMS jobs can also access data remotely via the CMS XRootD federation [7]. This setup offers an opportunity to explore the benefits of integrating data caches into the CMS ecosystem, optimizing task execution performance. Since major processing campaigns occur where the data is located, implementing CDN techniques becomes advantageous for CMS user analysis tasks, ensuring streamlined data accessibility and processing efficiency across the CMS network.

To support these initiatives, we have implemented an XCache service at PIC Tier-1 that serves data to both the PIC Tier-1 (Barcelona) and half of the compute nodes deployed at CIEMAT Tier-2 (Madrid) facilities. This service stores user data from remote sites, significantly reducing data access delays, improving CPU utilization, and potentially minimizing local storage needs. Comprehensive studies, performance measurements, and simulations have been conducted to evaluate the utility of the XCache service and determine optimal configuration settings. These efforts highlight our commitment to refining CMS data management and maximizing operational efficiency.

2.1 The CMS XCache Deployed in Spain

XRootD proxy cache (XCache [8]) is the preferred caching service for scientific data within WLCG [9], playing a pivotal role in the CDN-based infrastructure. XCache uses a physical cache to manage frequently accessed data. When a data request is made, the proxy checks the physical cache and delivers cached content promptly. If the data is not cached, the proxy retrieves it from the appropriate storage server via a hierarchy of re-directors, caching it for future requests to ensure optimal data accessibility and performance.

Following successful tests and validation, an XCache server was deployed at the PIC WLCG Tier-1 in Barcelona by 2021. This server has a capacity of 175 TiB, achieved with 6TB HDDs in RAID6 configuration. It is powered by 2xCPUs E5-2650L v3 (48 cores with HT enabled), 128 GB RAM, and an active-active 10 Gbps Network Interface Card (NIC), running XRootD 5.5.1. When the required input data for a job is not locally available, the CMS file opening fallback mechanism first engages with the XCache service. If the data is cached, it is served promptly; otherwise, it is fetched from a remote site via the CMS XRootD redirector infrastructure and then delivered from the cache to the compute node. XCache primarily handles file retrieval but can also be configured for read-ahead capabilities, retrieving data in blocks of 10x, each consisting of 50 kB, which is the current configuration set at PIC XCache.

The XCache service uses the Least Recently Used (LRU) deletion algorithm to efficiently manage outdated and unused data [10]. This algorithm organizes cached files by usage and timestamps, identifying long-unused files for removal. Deletion is triggered by watermarks representing specific occupancy thresholds. When occupancy exceeds the High-Watermark (HW) threshold of 95%, the algorithm deletes files until the Low-Watermark (LW) threshold of 90% is reached.

CMS uses local configuration files to establish rules for mapping Local File Names (LFNs) to Physical File Names (PFNs) at each site. These files define access protocols and data servers, accommodating both local and remote access through regular expressions. This configuration directs frequently accessed data to the XCache service while allowing the global redirection infrastructure to handle other data. Accurate knowledge of data popularity is crucial for optimal functionality. The PIC XCache service caches all CMS data types except for test files, intermediate files, and input pile-up sample files, which can reach up to 1 PB and are stored at FNAL Tier-1 and CERN Tier-0. The PIC XCache supports both the PIC Tier-1 and fifty percent of the compute nodes at CIEMAT Tier-2, collectively serving 4500 CPU cores. Daily, the XCache service currently manages access to approximately 5000 files, serving an average of 15 TB of data per day.

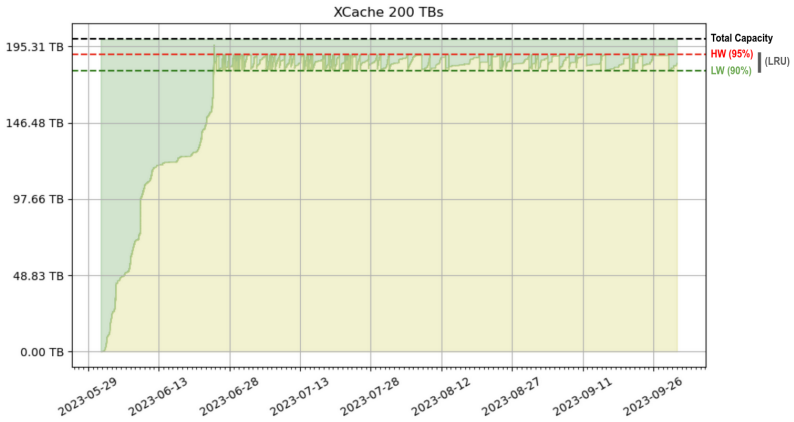

Figure 1: Simulation of a 200 TB XCache that caches all of the remote reads from CRAB jobs executed in PIC, CIEMAT and IFCA, in the period from June to October 2023.

By comparing the user's jobs at CIEMAT that use the XCache service, we have evaluated an average increase of ~ 10 % in the CPU efficiency for the user's submitted tasks [11]. This information is extracted from the user's CRAB [12] job logs. Since the effect is positive, we have analyzed all of the user jobs submitted to Spanish CMS facilities, for a big period of time, so we can simulate which could be the optimal cache size and characteristics deployed in PIC Tier-1, serving data to PIC Tier-1, CIEMAT Tier-2 and IFCA Tier-2 facilities, i.e. the sites in Spain that support CMS computational activities.

3 Optimal Dimensioning of a Single CMS XCache in Spain

Previous studies on cache algorithms and configurations for CMS [13] often lack the crucial information available in CRAB logs. To address this, our study utilizes CERN SWAN's Big Data infrastructure [14] and Dask's parallelization framework to efficiently process CRAB user logs. This approach enables us to construct a realistic model of data cache behavior, evaluate traditional caching algorithms for CMS regional XCaches, and determine optimal cache sizes. We analyzed CRAB jobs executed at Spanish sites over four months to emulate the impact of a cache serving the entire Spanish region. On average, 9.5k jobs were executed daily, with peak periods reaching 30k jobs. CIEMAT contributed 50% of these jobs, while PIC and IFCA (a Tier-2 in Santander) handled 34% and 16% of the workload, respectively.

3.1 Simulating XCache Implementation for Spanish CMS Tiers

Analyzing data access patterns from each Spanish CMS center is crucial for determining optimal cache size and network connectivity. We scrutinize all user job logs to identify whether data is accessed locally or remotely. While most files are fully downloaded, we also account for partial downloads, based on insights from the production PIC XCache service. Using this information, we simulate cache population to mirror real data access patterns from user jobs. To manage cache space efficiently, we apply the same LRU deletion algorithm and low/high watermarks as the production XCache, ensuring consistent and effective file management. Figure 1 shows a simulated XCache with a 200 TB capacity (black dashed line). The LRU deletion algorithm with 95%-90% watermarks (red and green lines) manages cached data deletions. The XCache disk reaches saturation in less than one month with the number of files created in the simulation.

An important aspect of a data cache, beyond its size, is its data import and export capabilities, which dictate the network connectivity required. Simulations have shown that the average daily import/export throughput exceeds the 10 Gbps capacity of the current NIC in the PIC production XCache service. Identifying potential bottlenecks through these simulations allows us to anticipate configurations that align with expected usage if all Spanish sites utilize the PIC XCache service. For a single XCache serving the entire Spanish region, a robust 25 Gbps NIC (active-active) is essential for seamless operation. An efficient XCache should prioritize exporting data to regional compute nodes rather than importing data from remote sites. Key metrics, such as the ratio of in/out average data rates and maximum in/out data rates, help determine the optimal XCache size for the region, ensuring its effectiveness for the CMS infrastructure.

Balancing the size of the cache involves retaining frequently accessed data while minimizing non-accessed files. Despite the LRU algorithm managing deletions, a cache that's too small fails to retain popular files, causing frequent re-population and unnecessary overhead. Conversely, an excessively large cache may hold outdated, unaccessed data through multiple LRU cycles. Characterizing the XCache is crucial to determine the most suitable size for deployment. The *Hit Rate* measure is pivotal in this context, quantifying cache effectiveness by the ratio of cache hits (accesses to files already present) to total accesses (hits and misses). A cache miss occurs when a file isn't found and must be cached anew. The *Hit Rate*, expressed as a percentage, can be calculated cumulatively as the data cache starts populating:

$$HitRate = \frac{hits}{hits + misses} = \frac{hits}{N_{accesses}} \tag{1}$$

3.2 Optimal XCache across Spain

Simulating different cache sizes is crucial for identifying the most efficient solution for the region. This process considers factors such as the *cumulative Hit Rate*, which measures the proportion of accesses to cached files compared to total accesses. Network considerations also play a key role in determining the optimal cache size, ensuring effective data transfer and accommodating user demands. By analyzing these factors together, we can identify the cache size that maximizes performance and resource utilization while minimizing overhead and inefficiencies.

Figure 2 (left) shows the *cumulative Hit Rate*, with the cache reaching around 60% for sizes exceeding 200 TB. Beyond this threshold, cache gains diminish significantly. Figure 2 (right) illustrates a balanced 3:1 ratio between outbound and inbound traffic for an optimally sized cache, indicating efficient data distribution and resource utilization. These findings suggest that a data cache of approximately 200 TB is sufficient for optimizing performance from the perspective of the *cumulative Hit Rate*. This size balances the region's caching needs without the diminishing returns of excessively large caches.

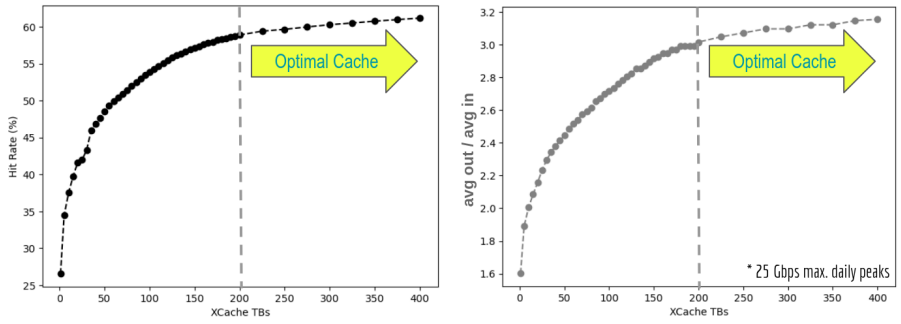

Figure 2: *Cumulative Hit Rate* (left) and ratio between outbound and inbound traffic at the XCache (right) for different simulated XCache sizes.

4 Conclusions

Our work demonstrates the advantages of utilizing the XCache service for optimized data access and resource utilization. XCache deployed in the PIC proves to efficiently serve data to all of the Spanish CMS sites, matching the effectiveness of local access at each Spanish CMS center, since all of these sites are located within a 10 ms round-trip time (RTT). The analysis of user job logs plays a crucial role in determining the optimal requirements for cache size and network connectivity. In our case, this analysis suggests the need for a cache size exceeding 200 TB, coupled with a NIC capacity of over 25 Gbps.

The integration of data caches holds promise beyond our immediate region, particularly in areas experiencing heightened remote data accesses by user tasks. These insights underscore the broader applicability and potential impact of data caching technologies in optimizing data workflows across diverse environments. In certain regions, remote reads surpass those observed in the Spanish sites. We have estimated that the traffic generated by remote reads from user jobs can reach up to 10 GB/s, a figure comparable to the global File Transfer Service (FTS) traffic generated by CMS worldwide. Consequently, the incorporation of data caches presents an opportunity to alleviate load on network resources and minimize overall traffic congestion. Data caches elsewhere could reduce the XRootD traffic generated by these user jobs' remote reads by (at least) a factor of 3.

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, DATA-2020-1-0039, and BES-2017-082665. 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. The deployment of the XCache service is financed by the Red Española de Supercomputación (RES) through the grant DATA-2020-1-0039.

References

[1] J. Albrecht, A. A. Alves, G. Amadio, G. Andronico, N. Anh-Ky, L. Aphecetche, J. Apostolakis, M. Asai, L. Atzori, M. Babik, G. Bagliesi, M. Bandieramonte, S. Banerjee, M. Barisits, L. A. Bauerdick, *A Roadmap for HEP Software and Computing R&D for the 2020s*, *Computing and Software for Big Science*, vol. 3, no. 7, pp. 1–39, Springer Science and Business Media LLC, Mar. 2019.

[2] J. Schovancova, S. Campana, X. Curull, M. Girone, I. Kadochnikov, G. McCance, *Understanding the Performance of a Prototype of a WLCG Data Lake for HL-LHC*, in *2018 IEEE 14th International Conference on e-Science (e-Science)*, pp. 332-333, IEEE Computer Society, 2018.

[3] Ian Bird, Simone Campana, Maria Girone, Xavier Espinal, Gavin McCance, *Architecture and prototype of a WLCG data lake for HL-LHC*, *EPJ Web Conf.*, vol. 214, pp. 04024, 2019.

[4] E. Martelli, S. Stancu, *LHCOPN and LHCONE: Status and Future Evolution*, *Journal of Physics: Conference Series*, vol. 664, no. 5, pp. 052025, Dec. 2015.

[5] Marian Babik, Shawn McKee, *Network Capabilities for the HL-LHC Era*, *EPJ Web Conf.*, vol. 245, pp. 07051, 2020.

[6] CMS Collaboration, *The CMS experiment at the CERN LHC*, *Journal of Instrumentation*, vol. 3, pp. S08004, 2008. Publisher: IOP Publishing.

[7] L. Bauerdick, K. Bloom, B. Bockelman, D. Bradley, S. Dasu, M. Ernst, R. Gardner, A. Hanushevsky, H. Ito, D. Lesny, P. McGuigan, S. McKee, O. Rind, H. Severini, I. Sfiligoi, M. Tadel, I. Vukotic, S. Williams, W. Yang, *Using XRootD to Federate Regional Storage*, *Journal of Physics: Conference Series*, vol. 396, no. 4, pp. 042009, 2012.

[8] L. Bauerdick, K. Bloom, B. Bockelman, D. Bradley, S. Dasu, Jeffrey Dost, Igor Sfiligoi, A. Tadel, Matevz Tadel, Frank Wuerthwein, A. Yagil, *XRootd, disk-based, caching proxy for optimization of data access, data placement and data replication*, *Journal of Physics: Conference Series*, vol. 513, pp. 042044, Jun. 2014.

[9] Teng Li, Robert Currie, Andrew Washbrook, *A data caching model for Tier 2 WLCG computing centres using XCache*, *EPJ Web Conf.*, vol. 214, pp. 04047, 2019. Section: T4 - Data handling.

[10] Alan Jay Smith, *Design of CPU Cache Memories*, in *Proc. IEEE TENCON*, 1987.

[11] C. Acosta-Silva, J. Casals, A. Delgado Peris, J. Flix Molina, J.M. Hernández, C. Morcillo Pérez, C. Pérez Dengra, A. Pérez-Calero Yzquierdo, F.J. Rodríguez Calonge and A. Sikora on behalf of the CMS Collaboration, *A case study of content delivery networks for the CMS experiment EPJ Web Conf.*, vol. 295, pp. 01006, 2024.

[12] Daniele Spiga, S. Lacaprara, William Bacchi, M. Cinquilli, Giuseppe Codispoti, M. Corvo, Alvise Dorigo, A. Fanfani, Federica Fanzago, Fabio Farina, O. Gutsche, Carlos Kavka, M. Merlo, Leonello Servoli, *CRAB: The CMS distributed analysis tool development and design*, *Nuclear Physics B - Proceedings Supplements*, vol. 177-178, pp. 267-268, Mar. 2008.

[13] Daniele Spiga, Diego Ciangottini, Mirco Tracolli, Tommaso Tedeschi, Daniele Cesini, Tommaso Boccali, Valentina Poggioni, Marco Baioletti, Valentin Y. Kuznetsov, *Smart Caching at CMS: applying AI to XCache edge services*, *EPJ Web Conf.*, vol. 245, pp. 04024, 2020.

[14] Danilo Piparo, Enric Tejedor, Pere Mato, Luca Mascetti, Jakub Moscicki, Massimo Lamanna, *SWAN: A service for interactive analysis in the cloud*, *Future Generation Computer Systems*, vol. 78, pp. 1071-1078, 2018.