ACAT 2024

Optimal XCache deployment for the CMS experiment in Spain

J. Flix [1,2], C. Pérez Dengra [1,2], A. Sikora [3], P. Serrano [3]

Port d'Informació Científica (PIC) [1], CIEMAT [2], UAB [3]

Motivation

The pivotal role of WLCG in handling data from LHC experiments underscores the necessity for its expansion and adaptation to meet HL-LHC demands. Global community provision of computational resources remains vital within budget constraints. Technological advancements offer relief, but ongoing R&D aims to manage future resources cost-effectively. The LHC community is exploring Content Delivery Network (CDN) techniques for optimized data access and resource utilization, deploying lightweight storage systems and enhancing task execution performance through efficient input data reading via content caching near end-users.

XCache deployment in Spain for the CMS experiment

Our previous studies **[1]** evaluate the benefits of implementing data cache solutions for the CMS experiment, focusing on Spanish compute facilities, indicating that user analysis tasks benefit most from CDN techniques. Consequently, a data cache has been introduced to gain deeper insights. The XCache at PIC Tier-1 has a capacity of 175 TiB, featuring a disk server with 6TB HDDs in RAID6, 2xCPUs E5-2650L v3 (48 cores), 128 GB RAM, and a 2x10 Gbps NIC. It currently serves data to both PIC Tier-1 and CIEMAT Tier-2, embedded with regional and CMS XRootD re-directors **[2]**.

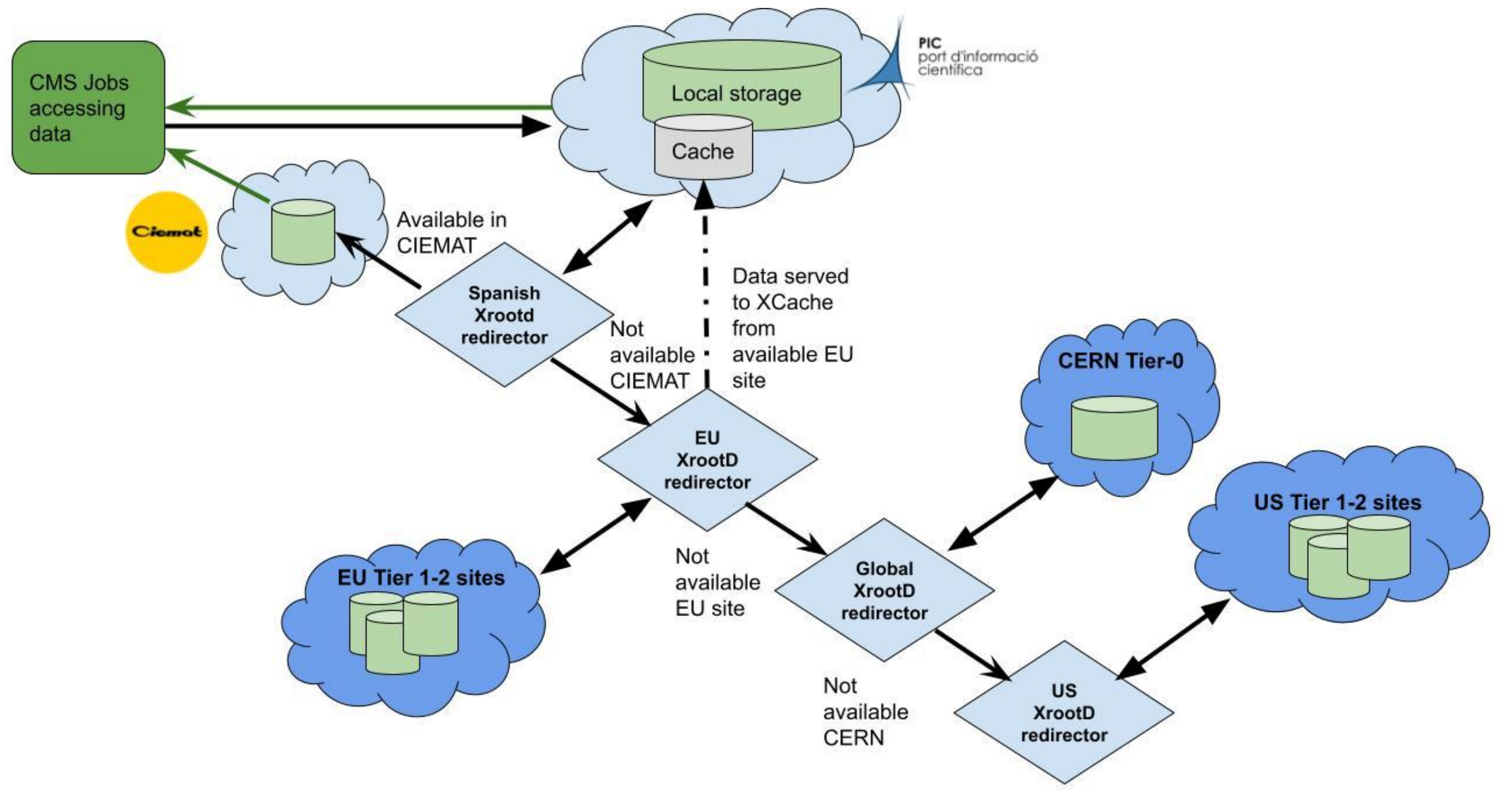


Figure 1: XCache configuration for remote input data reads of CMS jobs in PIC Tier-1 and CIEMAT CMS Tier-2

CMS CRAB Analysis logs have been scrutinized to showcase the efficiency improvements observed by end-users when utilizing the XCache service. Leveraging Big Data technologies like Hive-Spark and parallelization techniques such as Dask, several Terabytes of data from job logs executed in PIC and CIEMAT, accessing remote data or utilizing XCache, were analyzed. Comparing jobs that read remotely with those utilizing XCache, the relative increase in CPU efficiency for analysis tasks is estimated to be approximately 10% with the implementation of XCache **[3]**.

Remote reads from Analysis tasks executed in Spain

To model the impact of a cache serving data across Spain, approximately 1.2 million CRAB jobs spanning June to September 2023 were examined. Of these, around 50% were processed at CIEMAT Tier-2, while 34% and 16% were handled at PIC Tier-1 and IFCA Tier-2, respectively. Each CRAB job can access multiple input files, with an average of approximately 2.7 files per job during this period. The percentage of files accessed from local and remote storage elements (SEs) was computed daily for all Spanish sites. For PIC, CIEMAT, and IFCA, approximately 22%, 33%, and 77% of input files were retrieved from remote SEs, respectively. In total, around 3.1 million files were accessed during this period, with 1.1 million files accessed remotely.

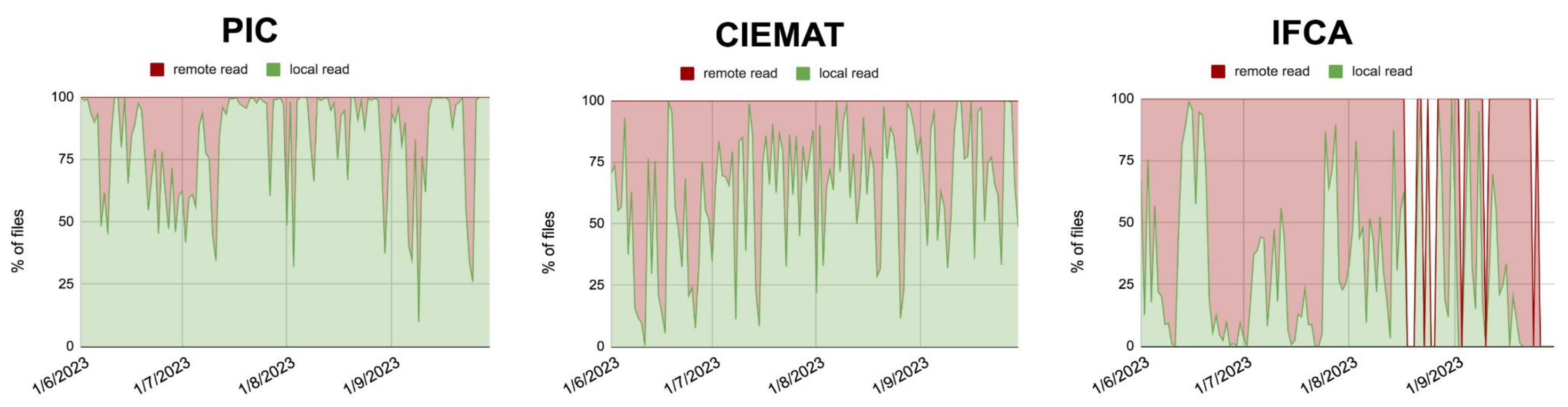
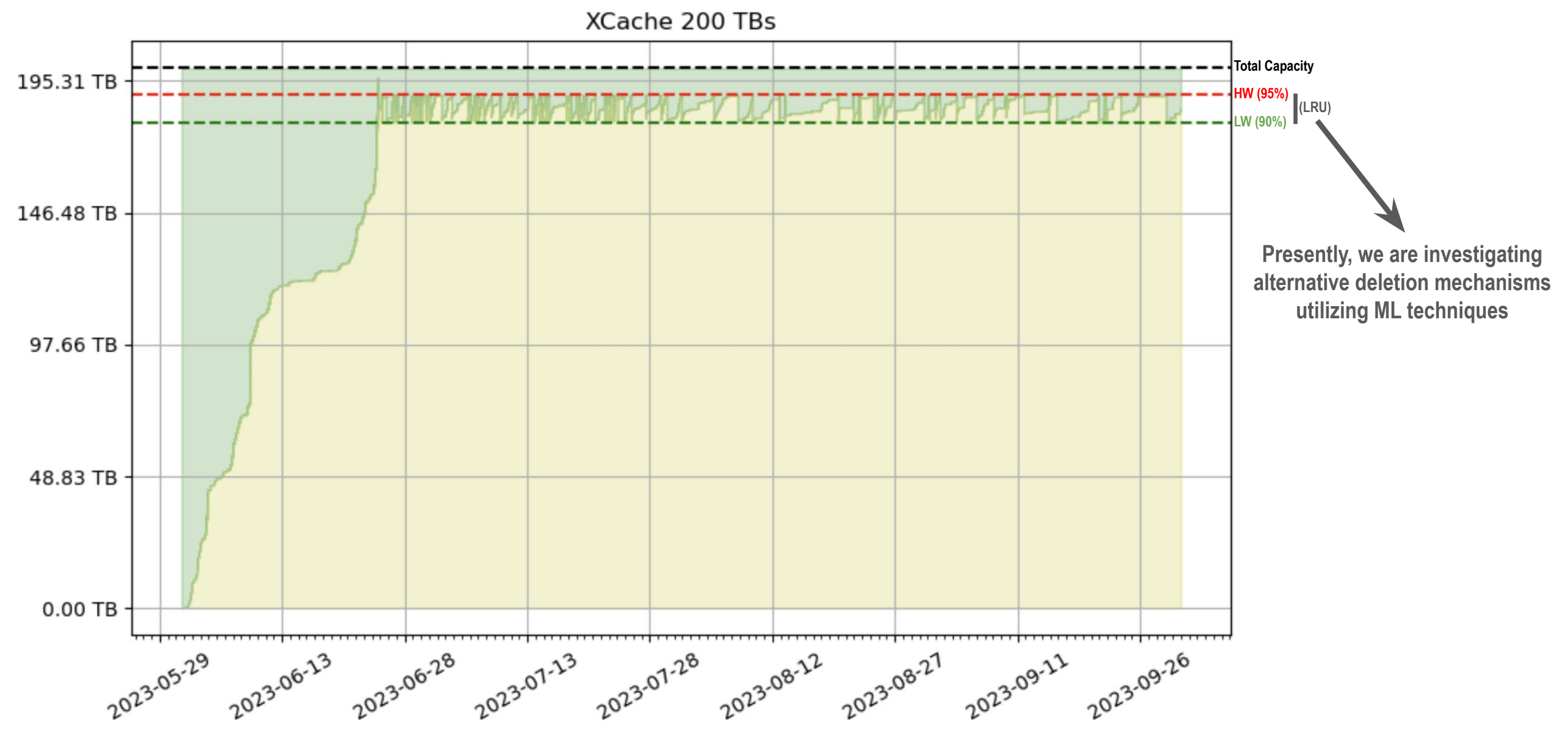


Figure 2: Percentage of input files that have been read from local SE (green) or from remote sites (red) in PIC, CIEMAT and IFCA

Simulating a cache for the region

The data access details help determine optimal requirements in cache size and network connectivity. Most files are fully downloaded, but we consider partial downloads using information from the production PIC XCache. Deletion from the cache is managed by the Least Recently Used (LRU) algorithm, triggered by watermarks representing occupancy thresholds. When occupancy surpasses the High-Watermark (HW) of 95%, the algorithm initiates file deletion until reaching the Low-Watermark (LW) of 90%, ensuring efficient space management.



Presently, we are investigating alternative deletion mechanisms utilizing ML techniques

Figure 3: Simulation of a 200 TB XCache that caches all of the remote reads from CRAB jobs executed in PIC, CIEMAT and IFCA

Optimal XCache across Spanish CMS Tiers

Simulating various cache sizes can identify the most efficient option for serving the region, determined by factors such as the cumulative Hit Rate (accesses to cached files over total accesses) and network considerations.

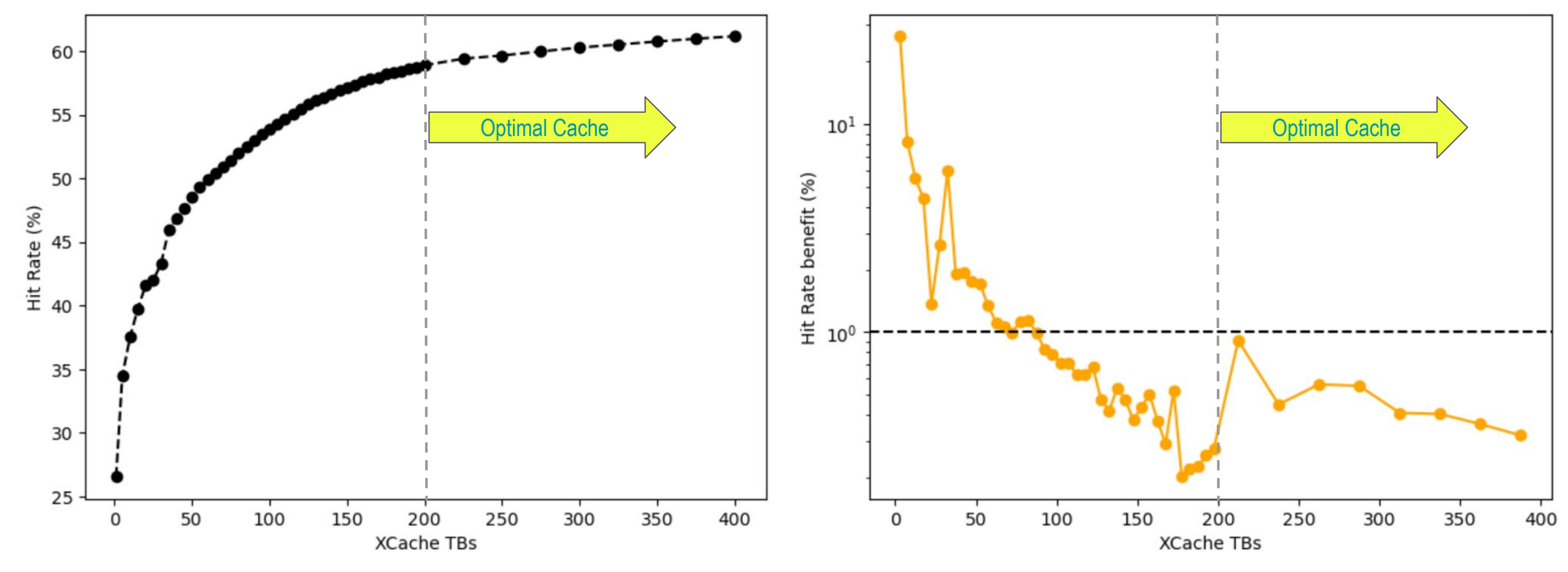


Figure 4: Cumulative hit rate (%) [left] and the percentage of hit rate gain or benefit when transitioning to a bigger data cache (a line is drawn at 1% level, for reference) [right]

To accommodate daily peaks in both data imports and exports, each simulated data cache requires a disk server with a 25 Gbps NIC. Opting for a 100 Gbps NIC would offer additional headroom, considering that the computed values represent daily averages and actual peaks during the day may exceed these estimates. An effectively dimensioned cache typically exhibits a 3:1 ratio between outbound and inbound traffic.

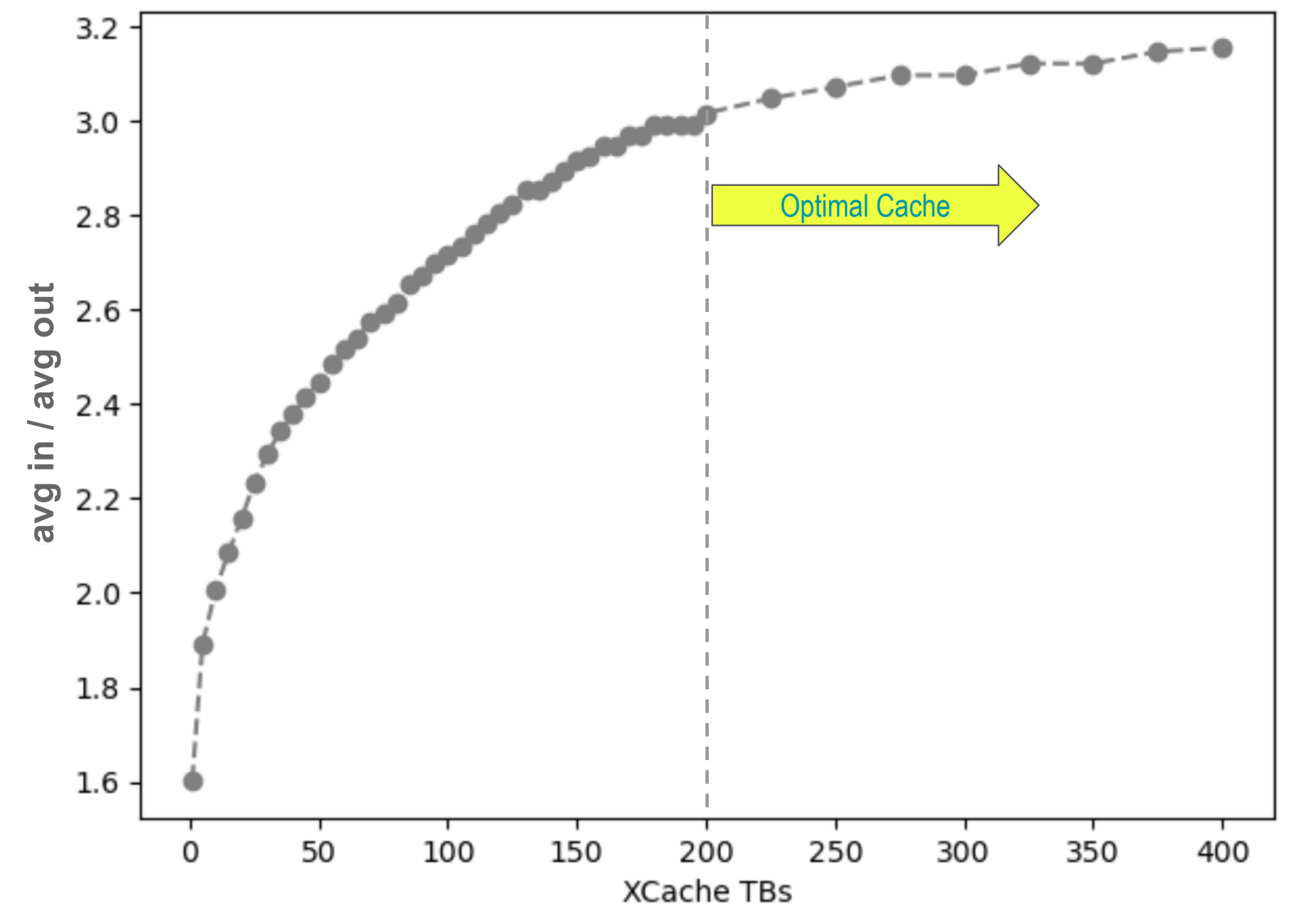


Figure 5: The ratio of average out to average in data rates, as a function of simulated XCache size

References

[1] *CMS data access and usage studies at PIC Tier-1 and CIEMAT Tier-2* [10.1051/epjconf/202024504028]

[2] *New storage solution for CMS experiment in Spain towards HL-LHC* [10.1088/1742-6596/2438/1/012053]

[3] *A case study of content delivery networks for the CMS experiment*, CHEP 2023 [cern.ch/chep2023]

Acknowledgements: 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, and the Red Española de Supercomputación (RES) through the grant DATA-2020-1-0039.

Contact: jflix@pic.es