



Content Delivery Network solutions for the CMS experiment: the evolution towards HL-LHC

J. Flix, C. Pérez, A. Sikora, P. Serrano

ISGC 2024
Taipei (Taiwan)
25-29 March 2024



The HL-LHC context

Currently expanding and adapting the World-Wide LHC Computing Grid (WLCG) to accommodate **increased data processing demands expected at the HL-LHC era**

- Emphasis on the need for **cost-effective solutions** to manage the growing volume of data
- **Proposal to consolidate storage** resources in fewer sites
- Introduction of **Content Delivery Network (CDN)** techniques as a strategy for optimized data access and resource utilization
- Focus on deploying lightweight storage systems (**data caches**) supporting traditional (Grid) and opportunistic (Cloud/HPC) compute resources
- **Boost task execution performance** by implementing efficient data caching mechanisms in close proximity to end users

The CMS context

Default behavior for **CMS jobs** is to process data at its location, but they also possess the **capability to access data remotely** through the **CMS XRootD federation**

This setup offers a distinctive opportunity to evaluate the **advantages of employing data caches** to optimize CMS task execution performance

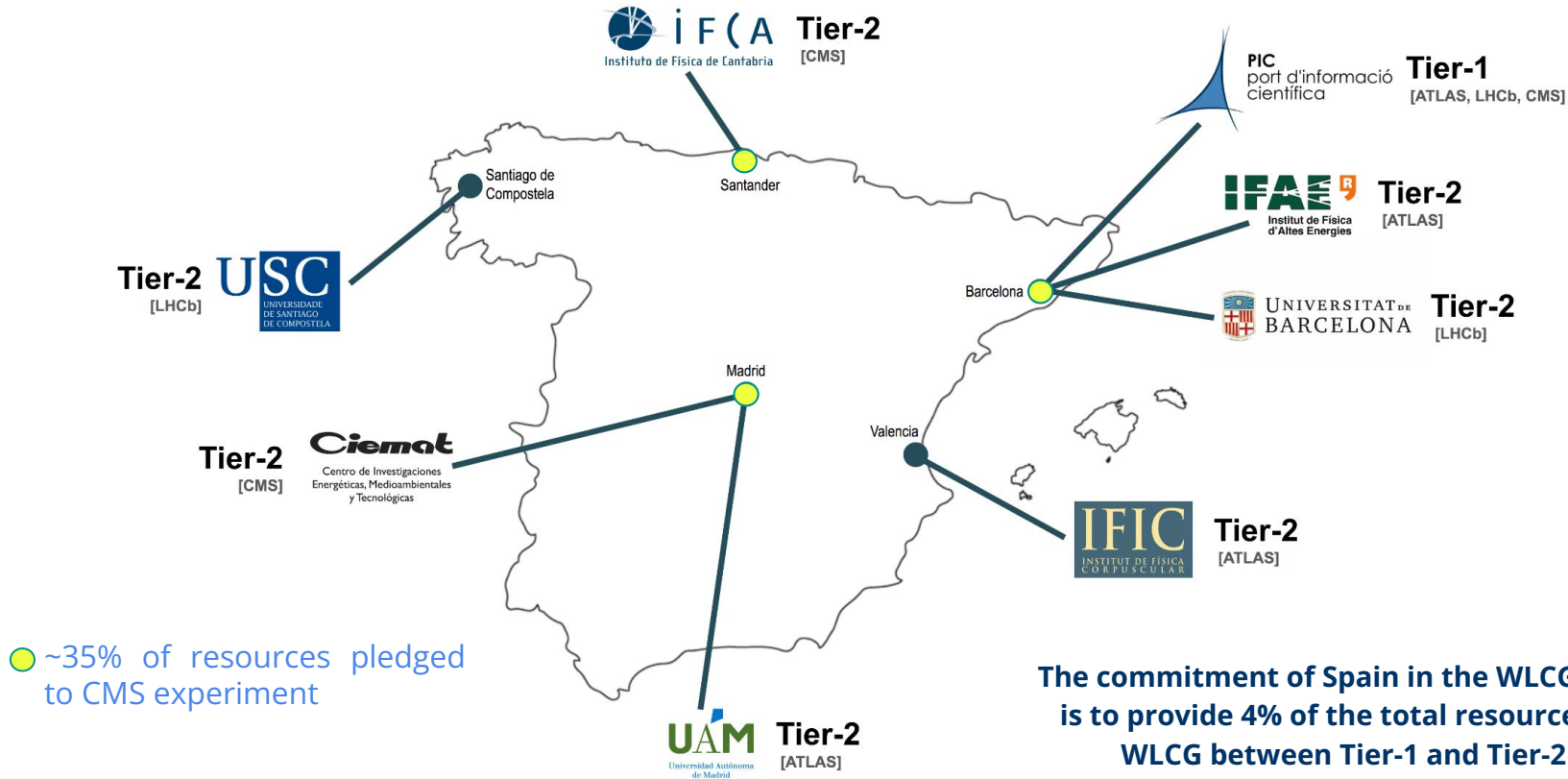
Major processing campaigns are run where data is placed, hence **CMS user analysis tasks benefit most from CDN techniques** [\[10.1051/epjconf/202024504028\]](https://doi.org/10.1051/epjconf/202024504028)

We have deployed an **XCache service** at **PIC Tier-1** and **CIEMAT Tier-2** to cache user's data which is read from remote sites

XCache helps **reducing data access latency, improving CPU efficiency** and potentially **reducing the storage** deployed in the region

Studies, performance **measurements**, and **simulations** have been performed to demonstrate the usefulness of the service and reach the best configuration

WLCG resources in Spain

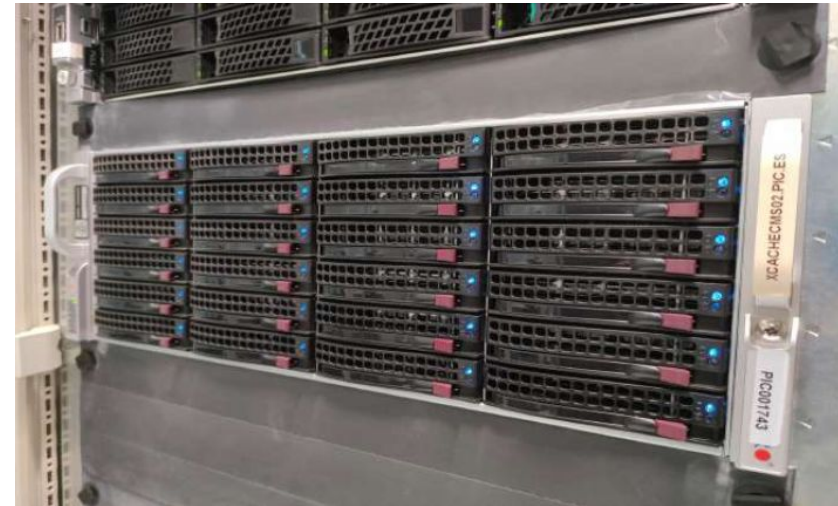


The XCache service deployed



The XCache at PIC Tier-1 has a capacity of **180 TB**, featuring a disk server with 6TB HDDs in RAID6, 2xCPU E5-2650L v3 (48 cores), 128 GB RAM, and a 2x10 Gbps NIC.

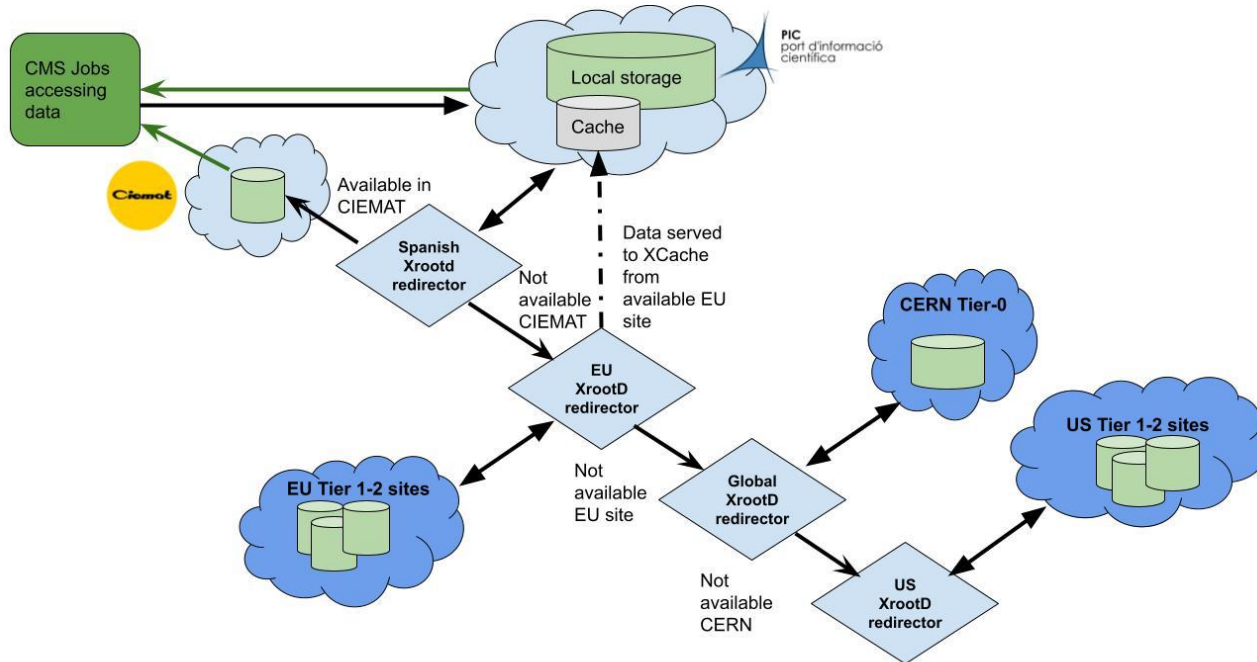
It currently **serves data to both PIC Tier-1 and CIEMAT Tier-2**



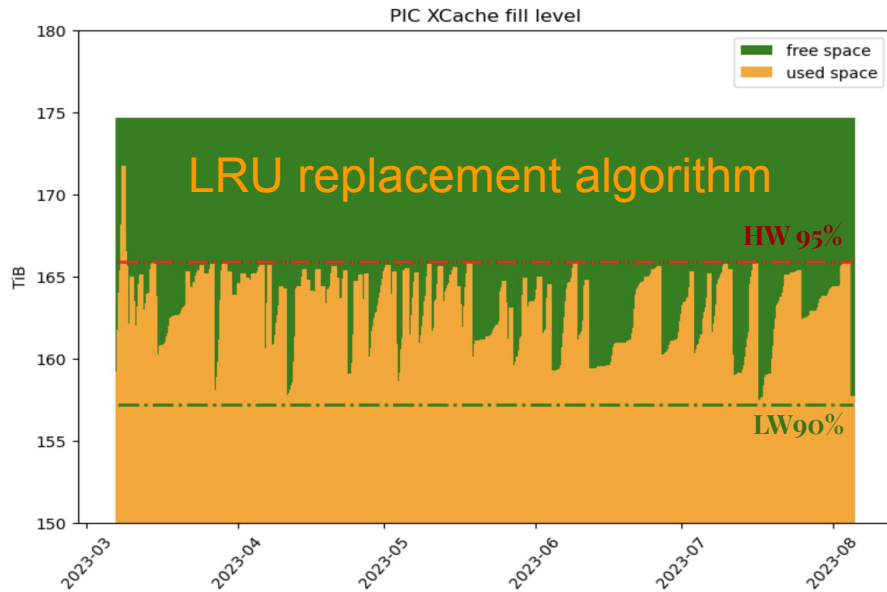
Deployment funded by the Spanish Supercomputing Network through project **DATA-2020-1-0039**

The XCache service deployed

The XCache at PIC Tier-1 is embedded with **regional** and **CMS XRootD re-directors**
[\[10.1088/1742-6596/2438/1/012053\]](https://doi.org/10.1088/1742-6596/2438/1/012053)

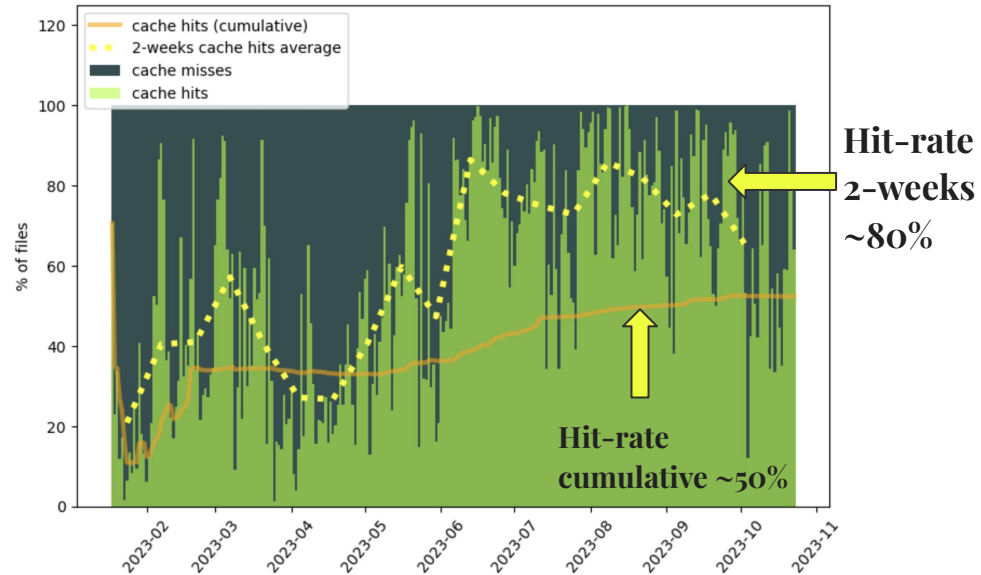


The XCache service deployed



In production, achieving good performance

How did we accomplish this?

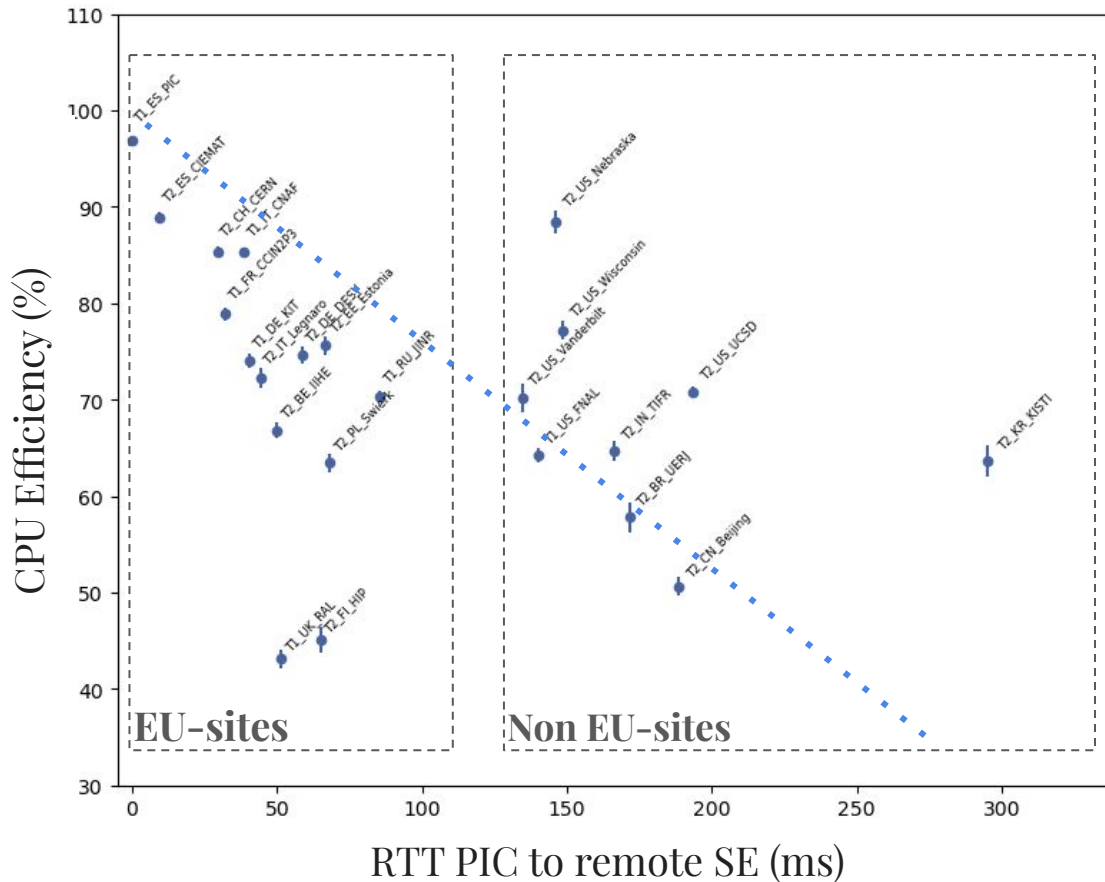


How severely do remote reads impact CMS user jobs?

CPUeff analysis benchmark job

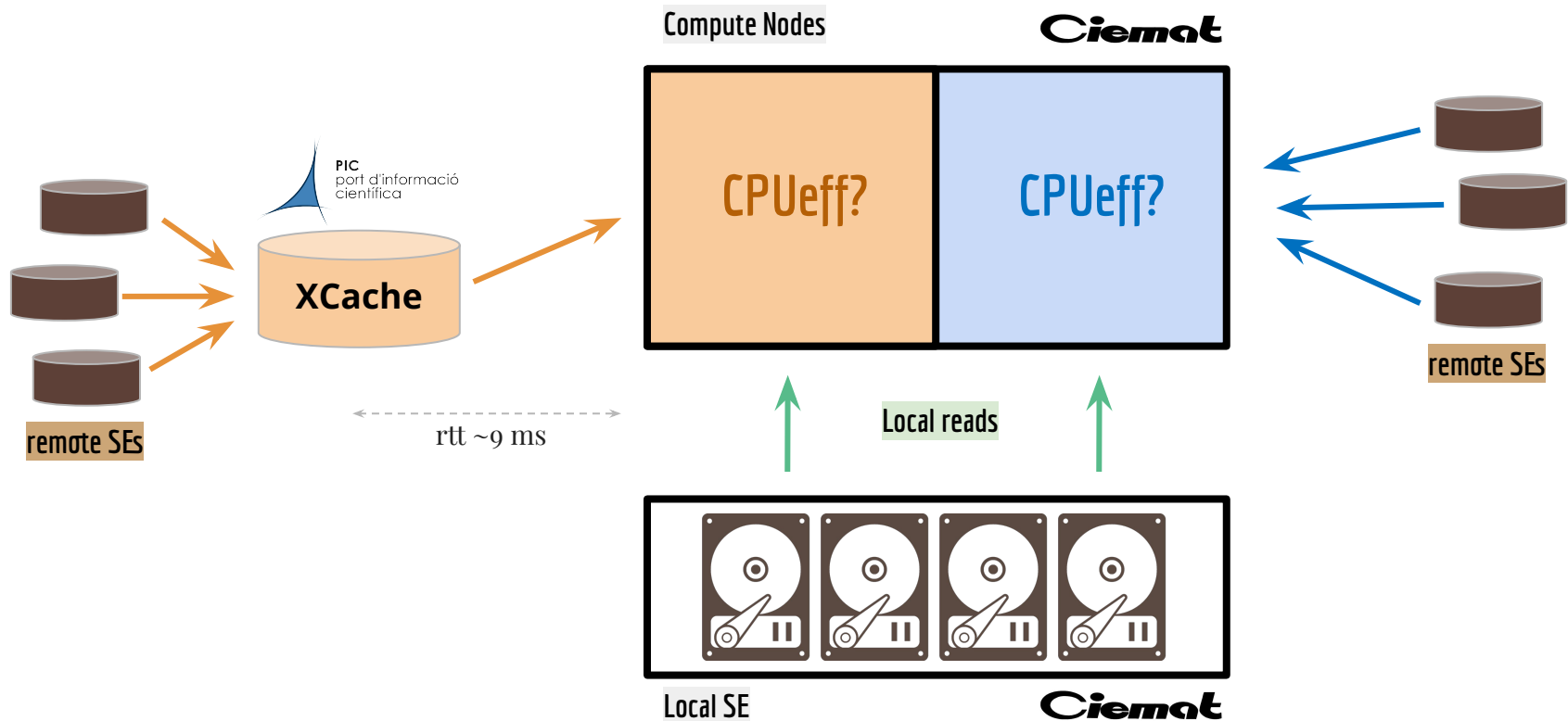


Execution of an Analysis **benchmark job** in PIC Tier-1 isolated compute node, reading MINIADO data from **local SE** or **remote SEs**

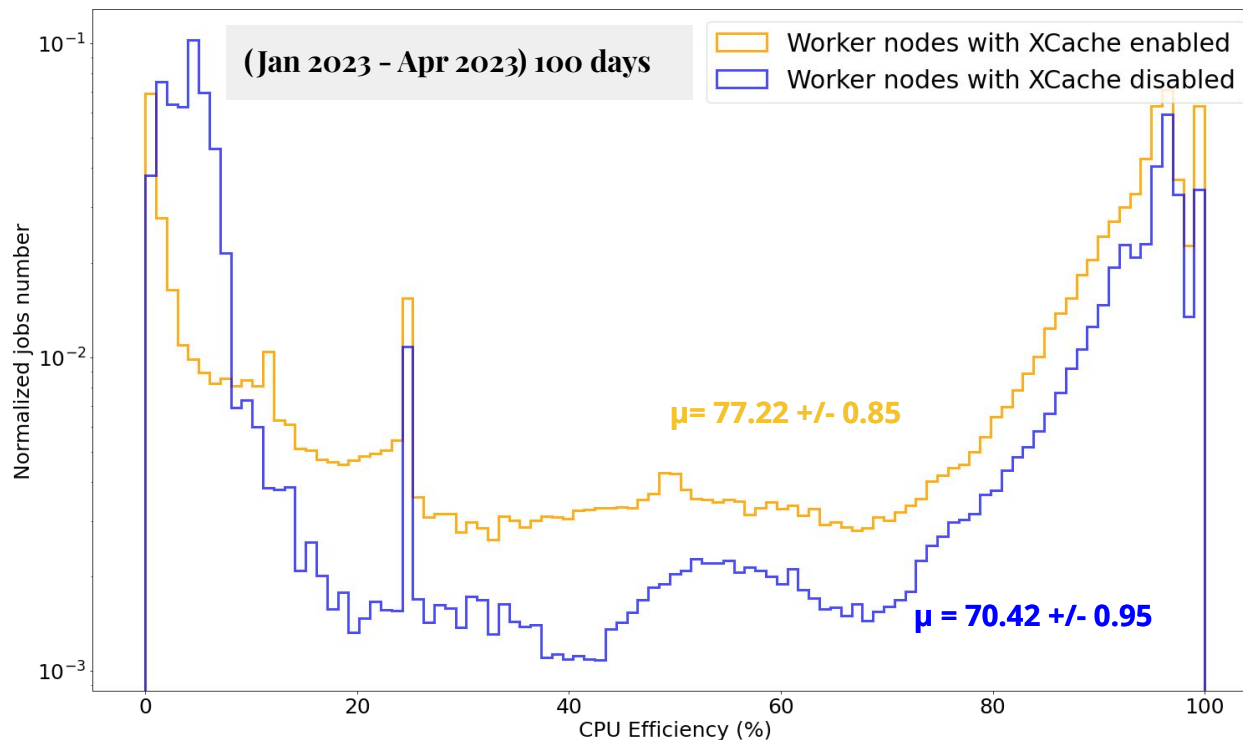


How severely do remote reads impact real CMS user jobs?

CPUeff real analysis jobs



CPUeff real analysis jobs



Overall improvement in CPU efficiency (~10%)

Insufficient CRAB monitoring → does not include
whether a read is remote or local

Local or remote reads?



CRAB logs contain local or remote reading information, they are stored in Ceph at CERN and accessible by HTTP through **cmsweb.cern.ch** (approx. last 3 months)

CERN's **SWAN** Big Data platform (Apache-Spark) gives access to the log's urls, that are downloaded and parsed at PIC using **Jupyter Notebooks + Dask**

→ Then we know **how many input files** have been read and **from where**, for each analysis user job

Size: 0.5MB/log → 7.5 TB/month for all of CMS sites → 0.3 TB/month for CMS Spain

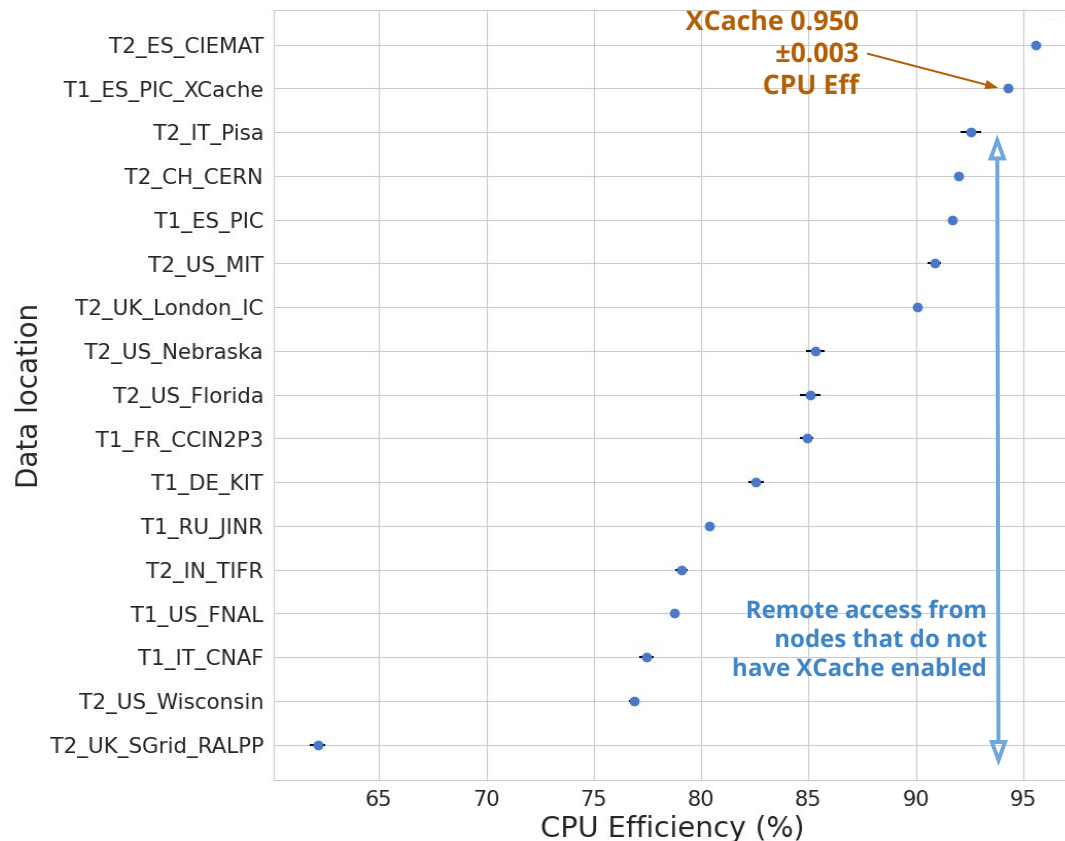
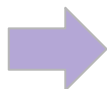


CPUeff real analysis jobs

The CPU efficiency of tasks reading from XCache is very close to those reading locally

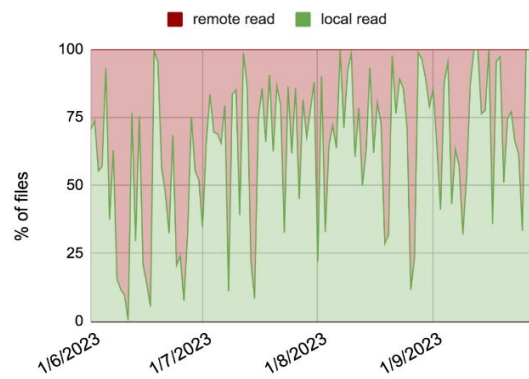
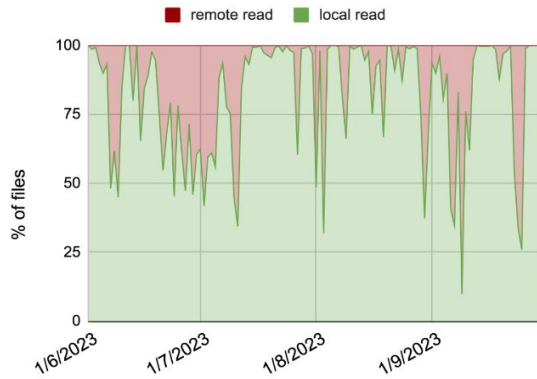
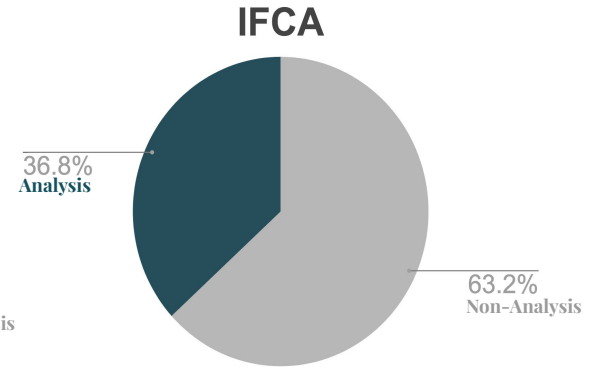
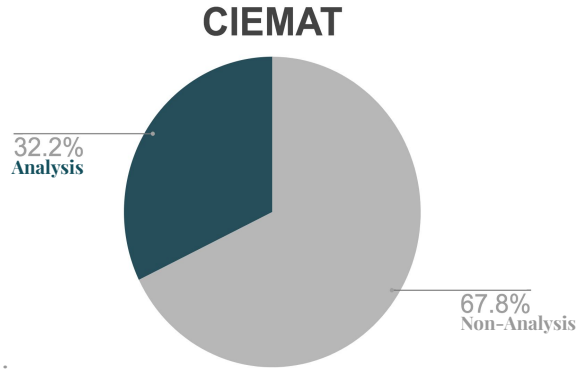
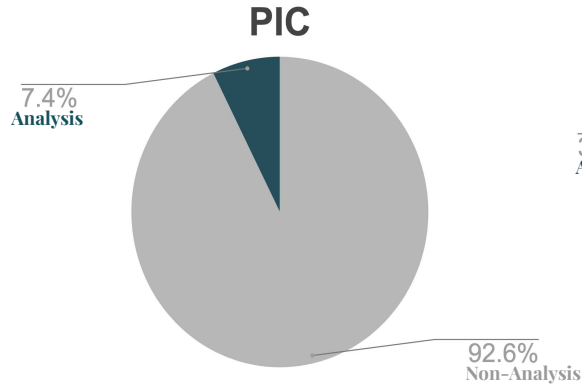
↳ Also better than those who read remotely

NOTE: These results correlate well with those obtained with the jobs benchmark studies



Which is the proper dimension of a single XCache to serve all the CMS sites in the Spanish region?

Remote reads: user's tasks in Spain



Many remote reads → potential to improve user analysis tasks performance

Simulating a cache for Spain



The **data access details** from all of the Spanish centers play a crucial role in determining the **optimal requirements for cache size** and **network connectivity**

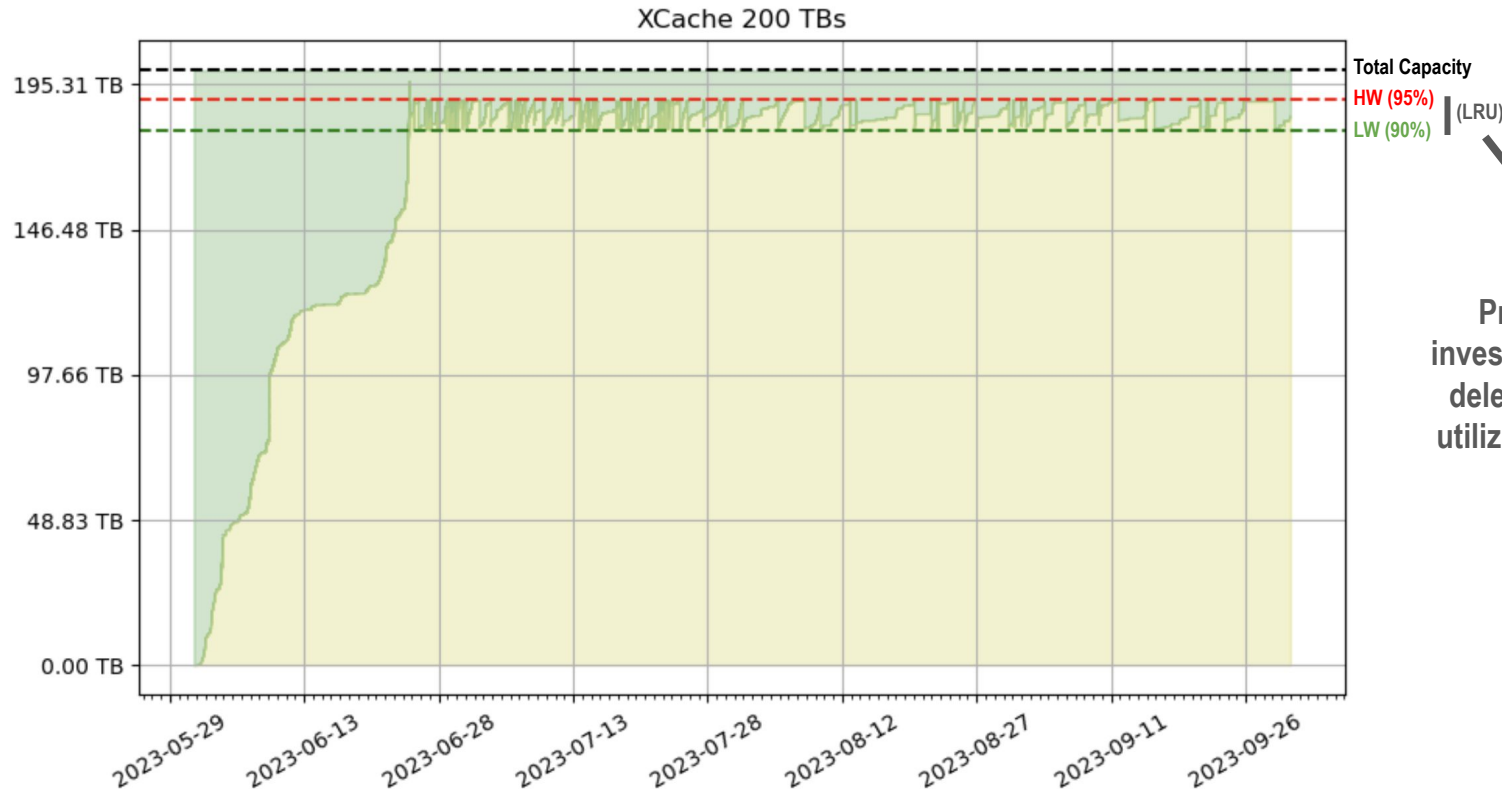
We explore all of the **user's job logs** for **local** or **remote** reading information

While most files are fully downloaded, **partial downloads** are considered based on insights from the production PIC XCache

We **emulate cache system population** based on these remote data accesses from production user's jobs

Deletion from the cache follows the **Least Recently Used (LRU) algorithm**: when occupancy exceeds 95% (High-Watermark - HW), file deletion is triggered until reaching the Low-Watermark (LW) of 90% for efficient space management

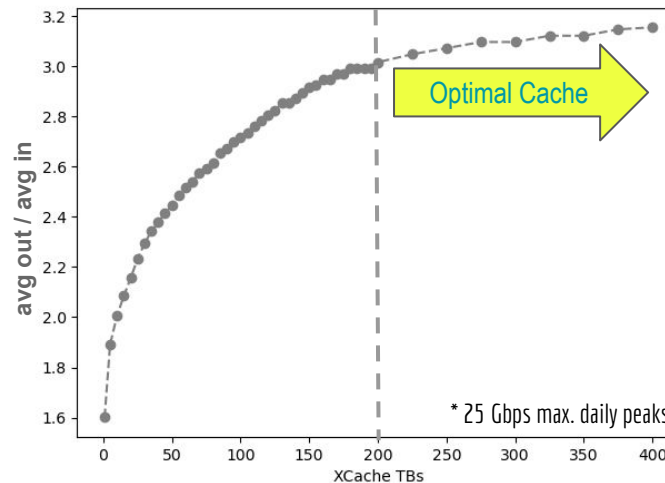
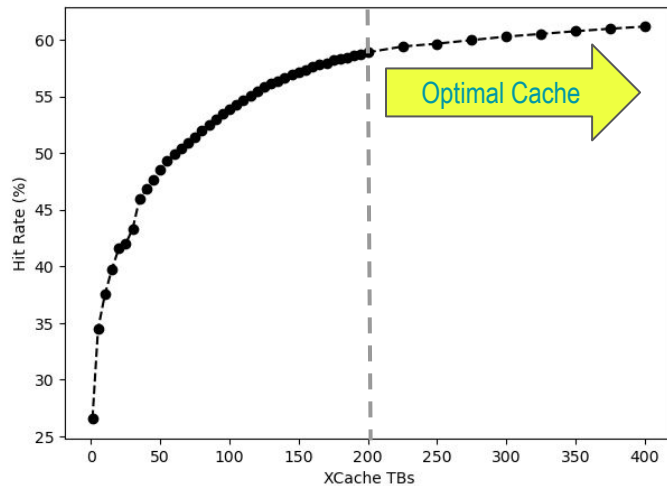
Simulating a cache for Spain: example



Presently, we are
investigating alternative
deletion mechanisms
utilizing ML techniques

Optimal XCache across Spain

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**



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


An effectively dimensioned cache typically exhibits a 3:1 ratio between outbound and inbound traffic

Conclusions

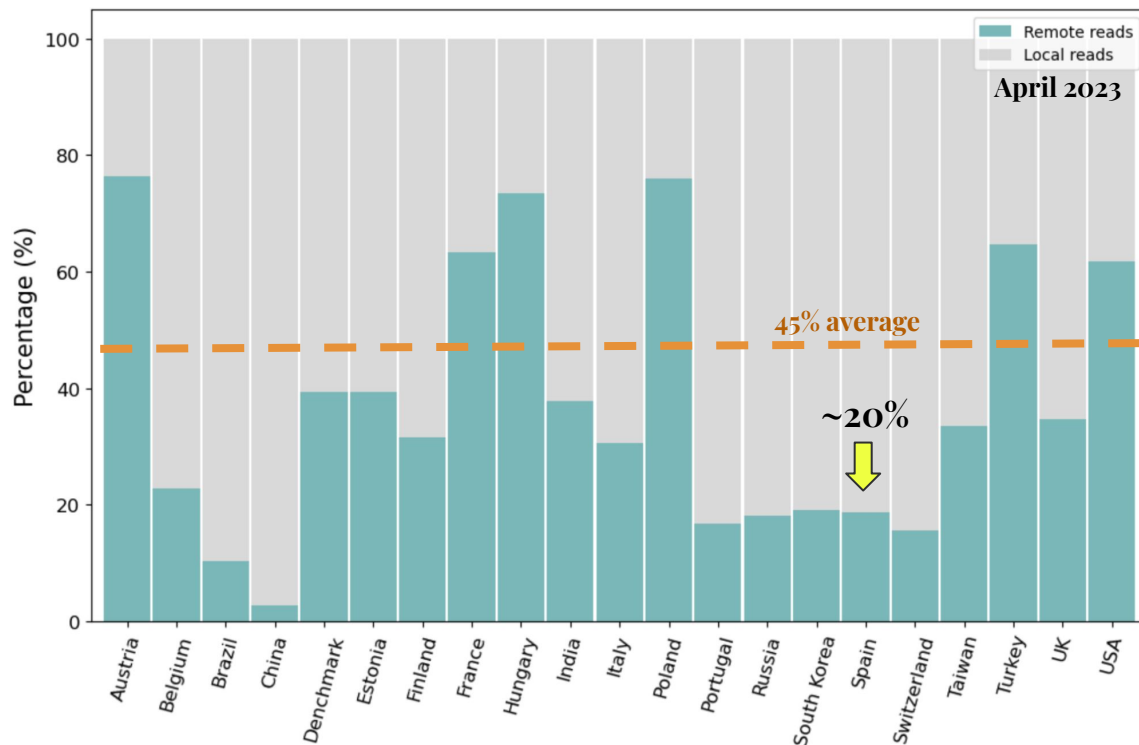
Our work exhibits the **advantages of the XCache service** for optimized data access and resource utilization

XCache deployed in the PIC **can efficiently serve data** just as effectively as if it were read locally at each of the Spanish CMS centers (sites within 10 ms RTT)

Analyzing user's job logs is instrumental to determine the **optimal requirements for cache size** and **network connectivity** (in our case: >200 TBs, with >25 Gbps NIC)

The utilization of data caches could have a **significant impact in other regions**, given their increased remote data accesses by user tasks 

Conclusions



The volume of data accessed remotely this month can be converted to GB/s

26 PB \rightsquigarrow 10 GB/s

Remote CRAB reads

Similar to the FTS traffic we had in April 2023!

Conclusions



Our work exhibits the **advantages of the XCache service** for optimized data access and resource utilization

XCache deployed in the PIC **can efficiently serve data** just as effectively as if it were read locally at each of the Spanish CMS centers (sites within 10 ms RTT)

Analyzing user's job logs is instrumental to determine the **optimal requirements for cache size** and **network connectivity** (in our case: >200 TBs, with >25 Gbps NIC)

The utilization of data caches could have a **significant impact in other regions**, given their increased remote data accesses by user tasks

In addition to the improved performance for CMS analysis tasks, introducing data **caches** elsewhere could **reduce the XRootD traffic generated by these user jobs'** remote reads by (at least) a factor of 3

謝謝

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