Norfolk, Virginia, USA • **May 8-12, 2023**

CHEP 2023

Computing in High Energy & Nuclear Physics

A case study of content delivery networks for the CMS experiment

C.Perez Dengra

J. Flix, A. Sikora, J. Casals, C. Acosta-Silva, C. M. Morcillo Perez, A. Pérez-Calero Yzquierdo, A. Delgado Peris, J. M. Hernández, F. J. Rodriguez Calonge for the CMS collaboration

PIC port d'informació científica

CMS

IFAE

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

Ciemat Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas

cfp CIEMAT física de partículas

UAB Universitat Autònoma de Barcelona

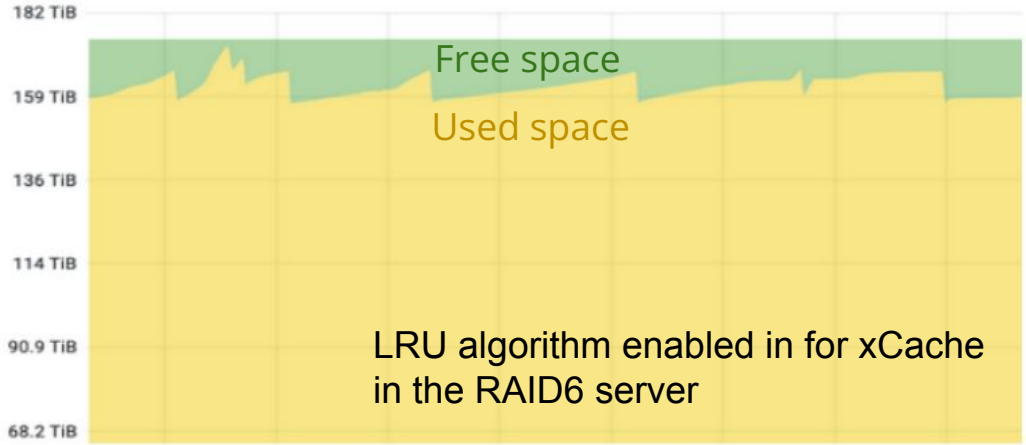
Context

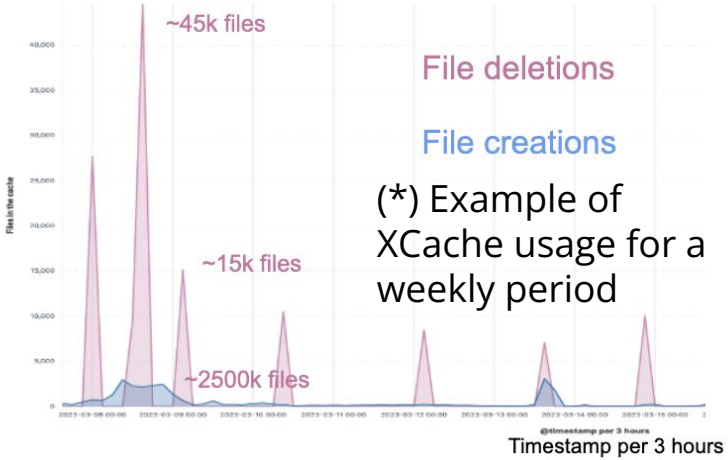
- **CMS jobs are sent to where the data is by default** but**, occasionally,** have the **capability to read data remotely** using the **CMS XRootD federation** (overflow to close sites and files opened in fallback).
- We have been **exploring the XCache** service at PIC Tier-1 and CIEMAT Tier-2 to cache data which is read from remote sites.
- XCache potentially helps **reducing data access latency**, **improving CPU efficiency** and **reducing the storage** deployed in the region
- We have deployed XCache service in the region, and dedicated **studies and performance measurements** have been performed to demonstrate the usefulness of the service and reach the best possible configuration.

XCache deployed at PIC and CIEMAT for CMS

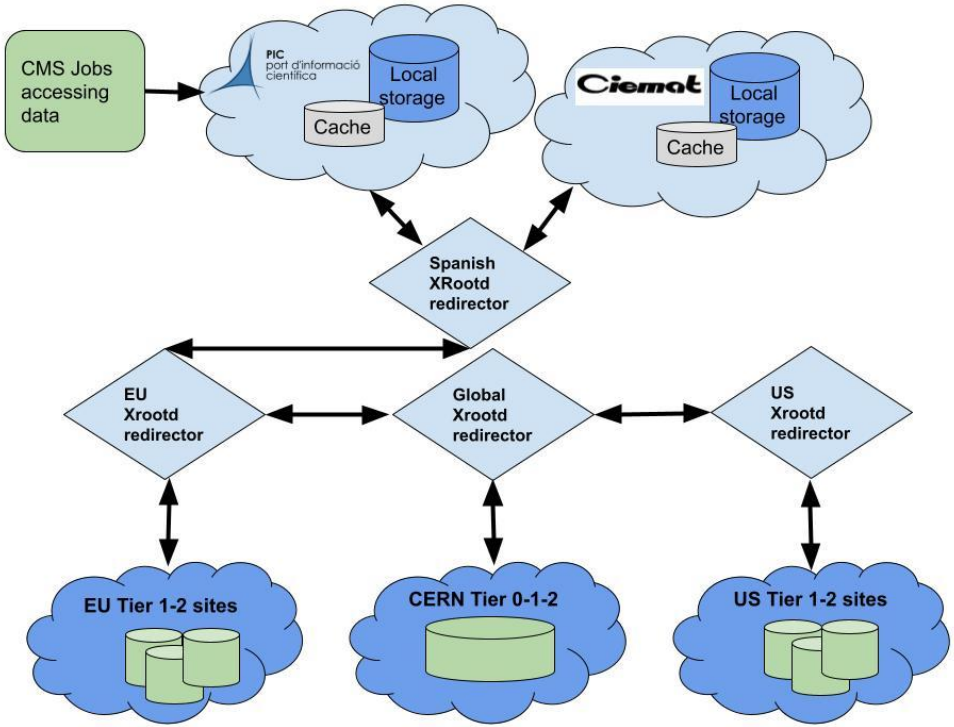
Currently, a **XCache service at PIC (180TB)** is serving data for **both PIC and CIEMAT** worker nodes.
PIC and CIEMAT rtt is ~9 ms → this deployment is sufficient to cater to the needs of both sites.

Deployment	Storage	XrootD version (OSG)	CMS data tiers to cache	Accessibility
Current (single cache for both sites)	T1_ES_PIC: 180 TB (RAID)	5.5.1-1	All (except pileup and unmerged data)	All worker nodes + half of the worker nodes at CIEMAT





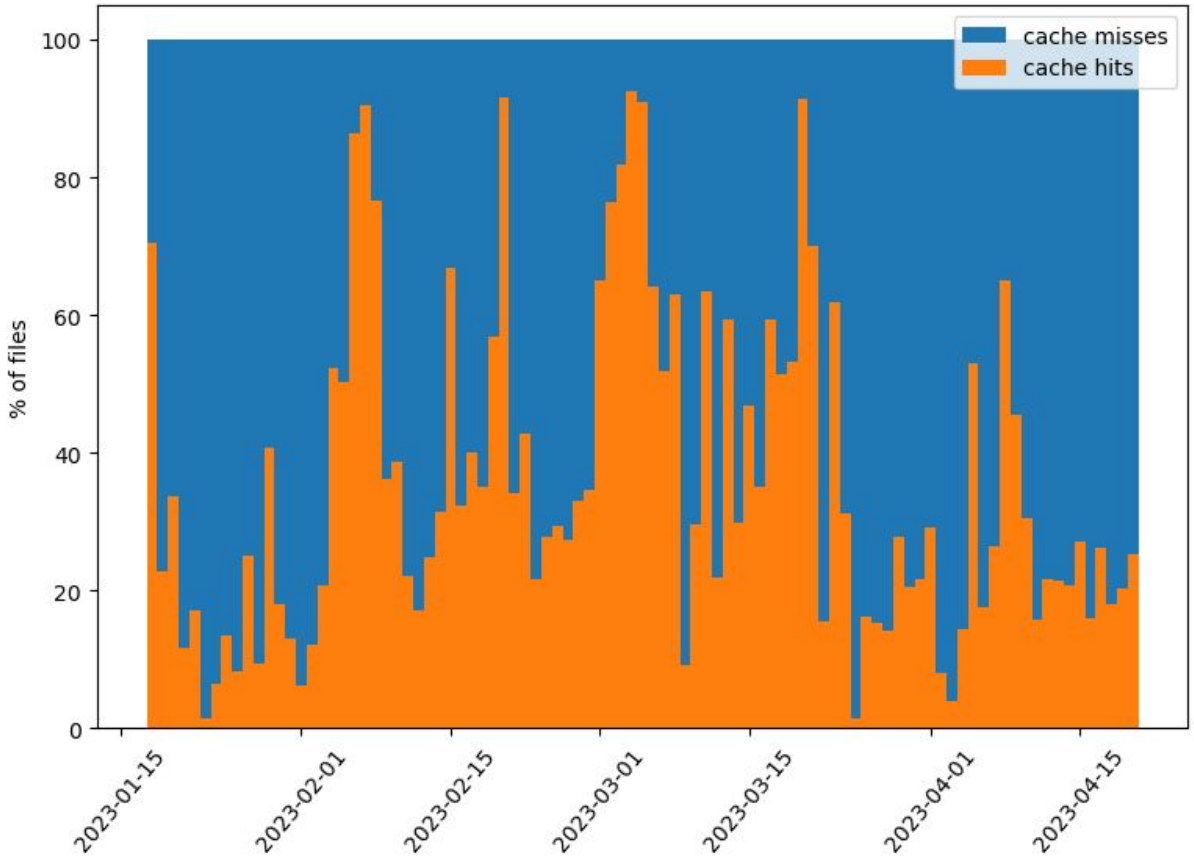
XCache configuration at PIC



Daily cache requests

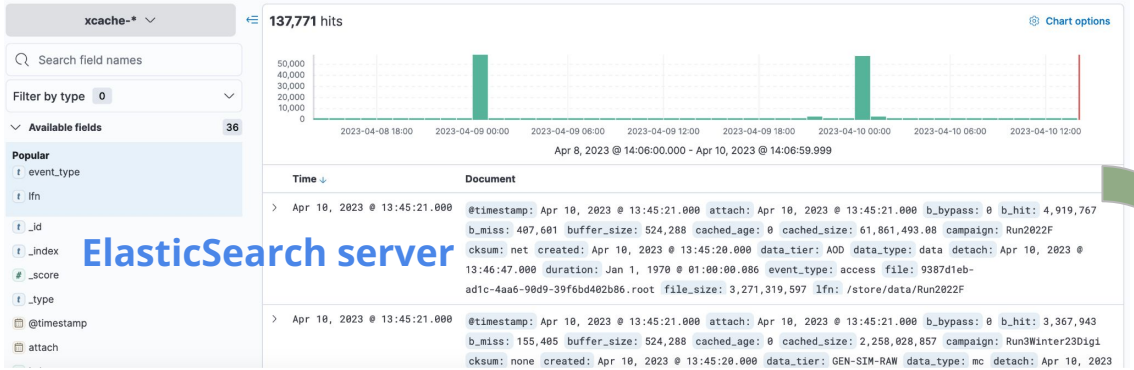
Since the commissioning of the centralized XCache at PIC. **About ~5000 daily files are daily created in average and ~2000 accesses received** (**average file size of 2.7GB**) from jobs executed at both sites requesting remote data.

The cache **hit-rate can fluctuate between periods of high and low performance,** depending on the **type of jobs executed.** However, there is always potential for improvement in this aspect.



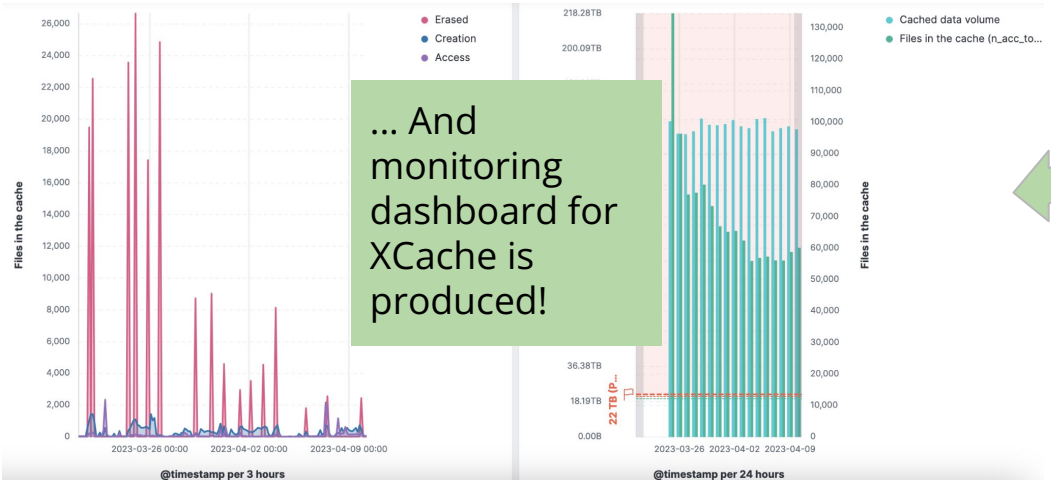
Local monitoring of XCache in PIC and CIEMAT

Each file downloaded produces a .cinfo file → monit binary files



ElasticSearch server

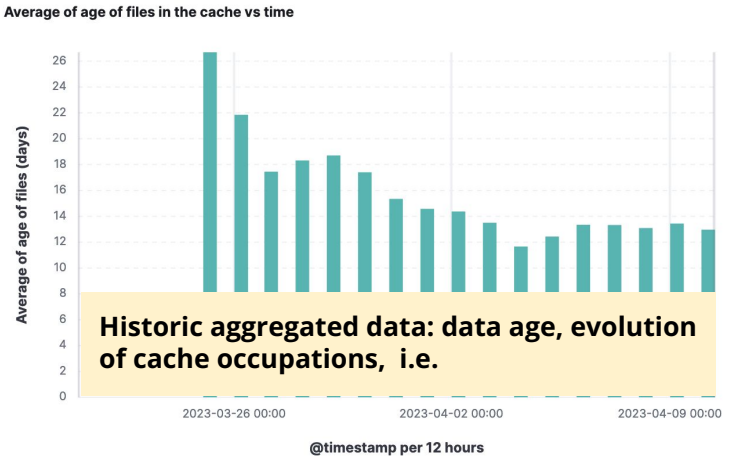
Events of **'creation', 'accesses' and 'deletions'** are processed …



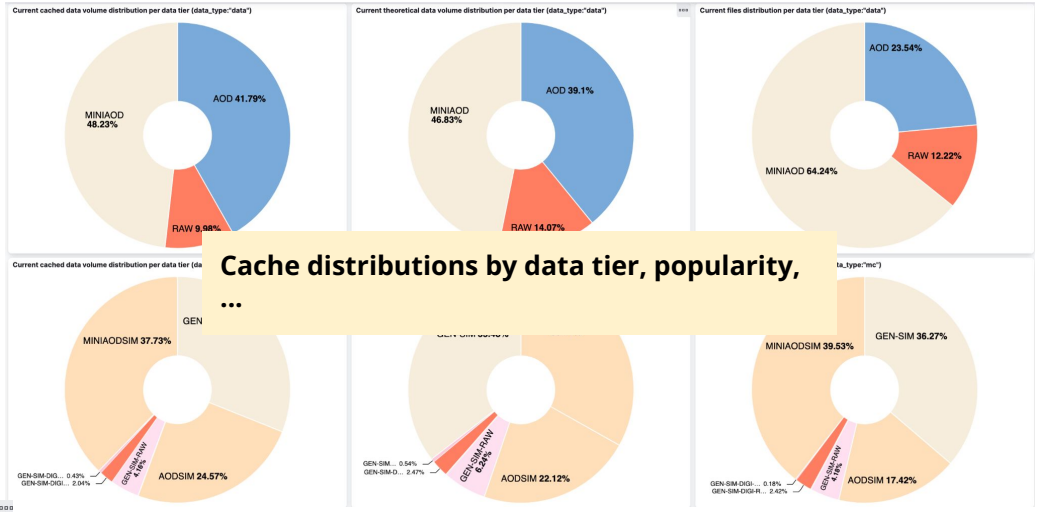
… And monitoring dashboard for XCache is produced!

Local monitoring of XCache in PIC and CIEMAT

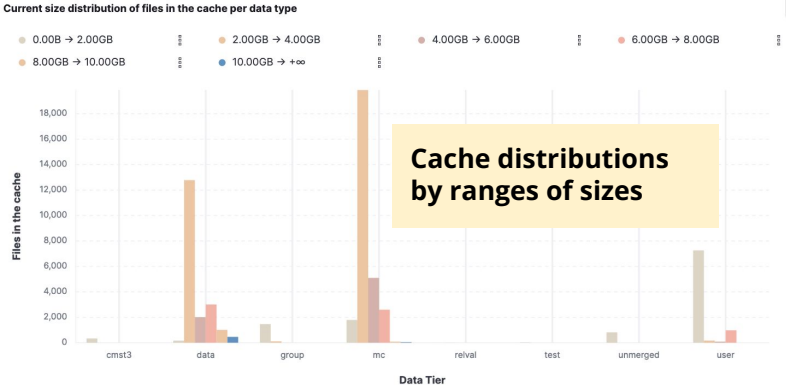
PIC
port d'informació
científica



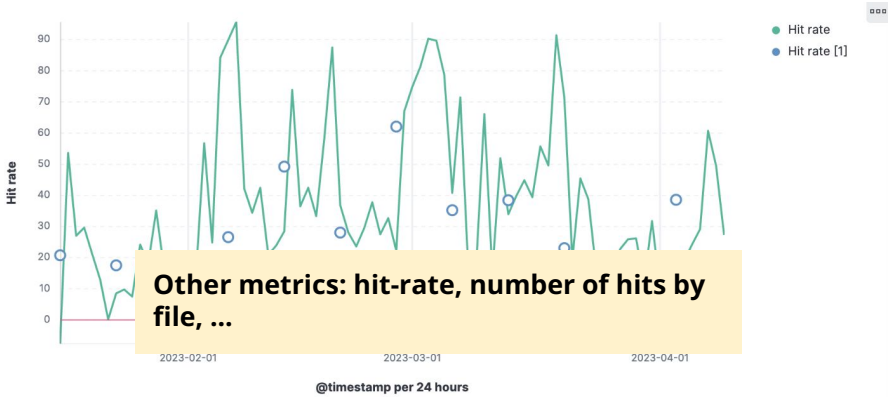
Historic aggregated data: data age, evolution of cache occupations, i.e.



Cache distributions by data tier, popularity, ...



Cache distributions by ranges of sizes



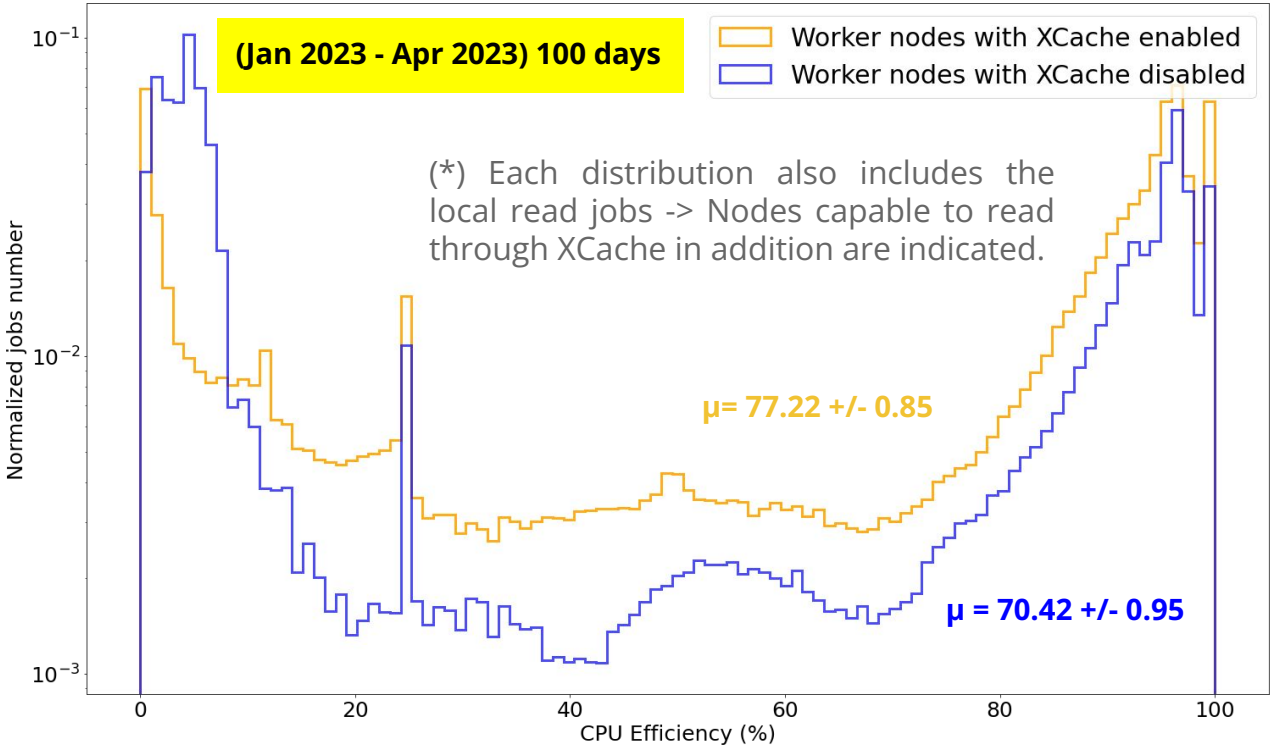
Other metrics: hit-rate, number of hits by file, ...

Analysis jobs at CIEMAT accessing data from different locations

Half of the worker nodes at CIEMAT have enabled the access to PIC's XCache.

CMS Remote Analysis Builder (**CRAB**) is a tool for distributed **data analysis in CMS.** Those are **the selected analysis jobs in the study.**

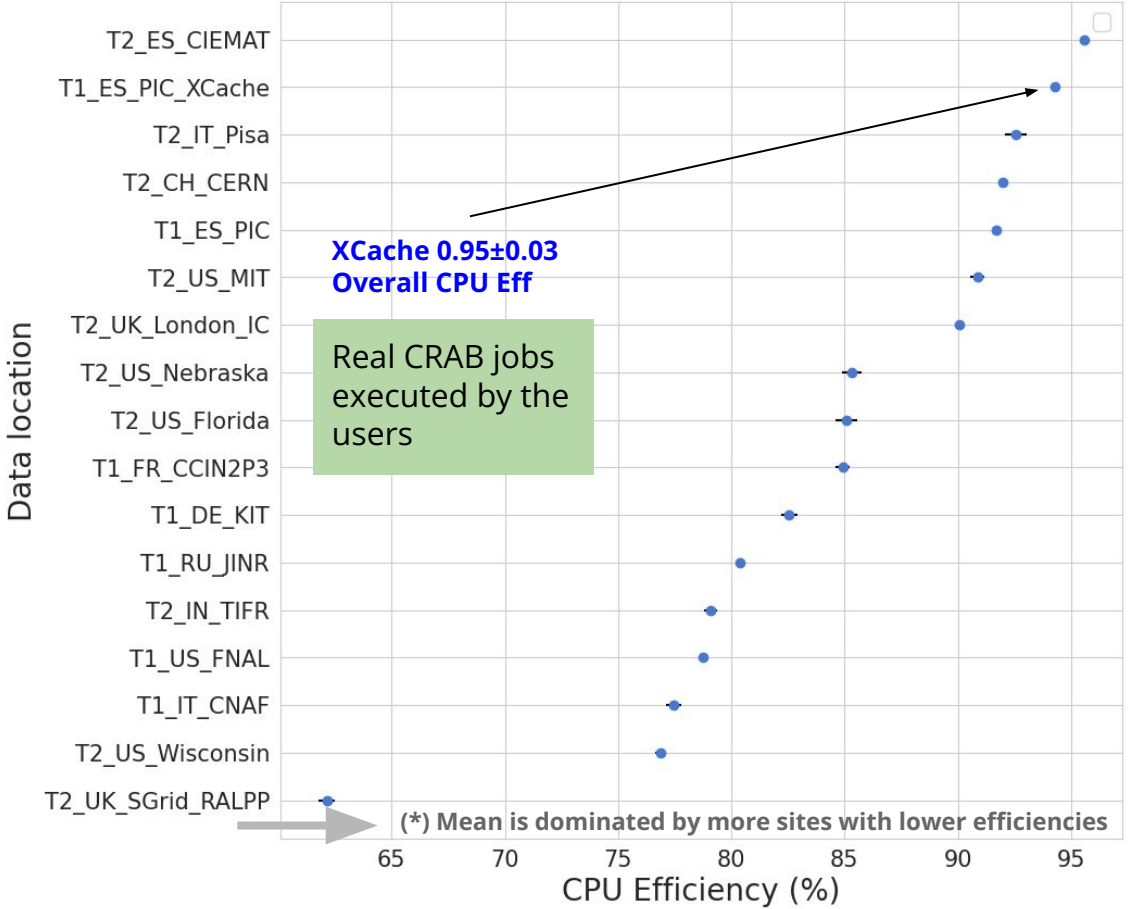
We have conducted a study to **compare the degradation of CPU efficiency** in CRAB jobs when executed at CIEMAT worker nodes **when caching is enabled or disabled.**



Analysis jobs at CIEMAT accessing remote data and PIC's XCache

We compare the **CPU efficiency of the standard CRAB jobs** executed at CIEMAT worker nodes by CMS users while **reading data from different CMS sites** with **those reading through XCache**.

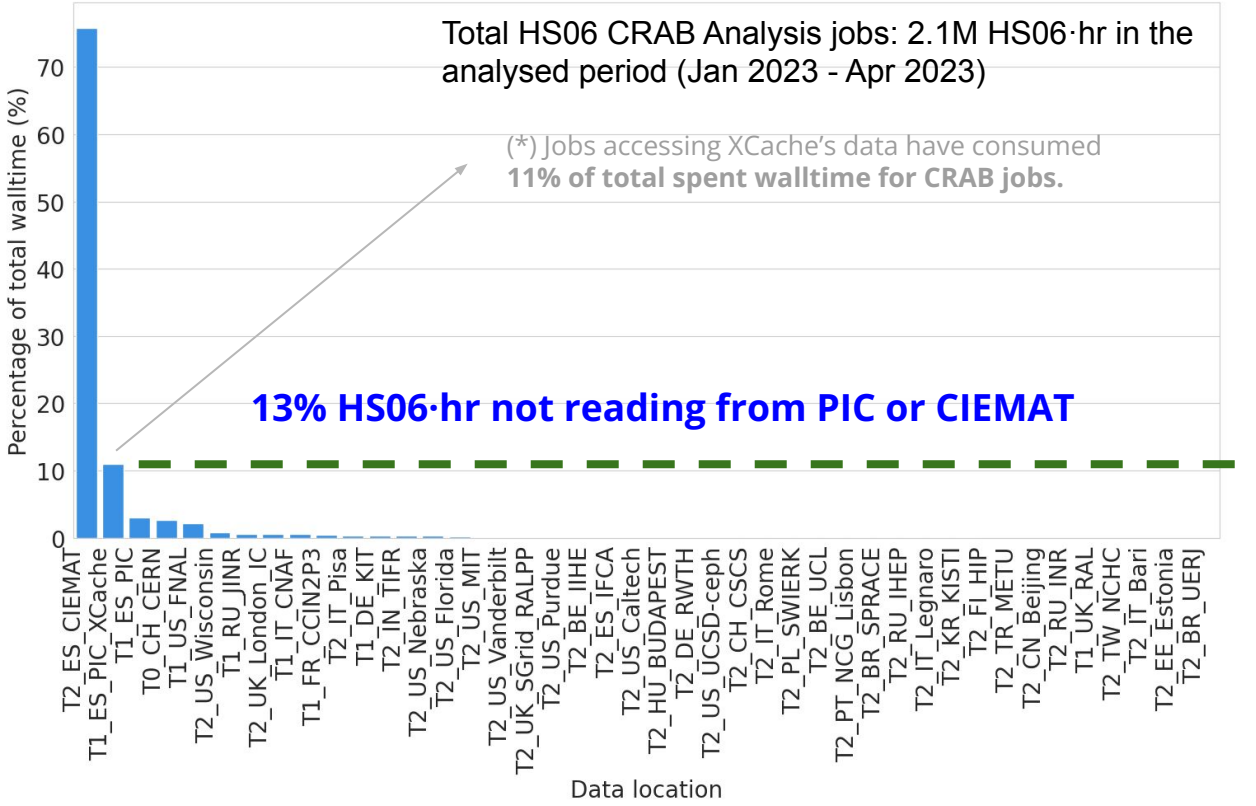
If **data is already in the cache**, the average CPU efficiency is **very close to that when reading data locally from CIEMAT**.



HS06·hr consumed by CRAB jobs at CIEMAT

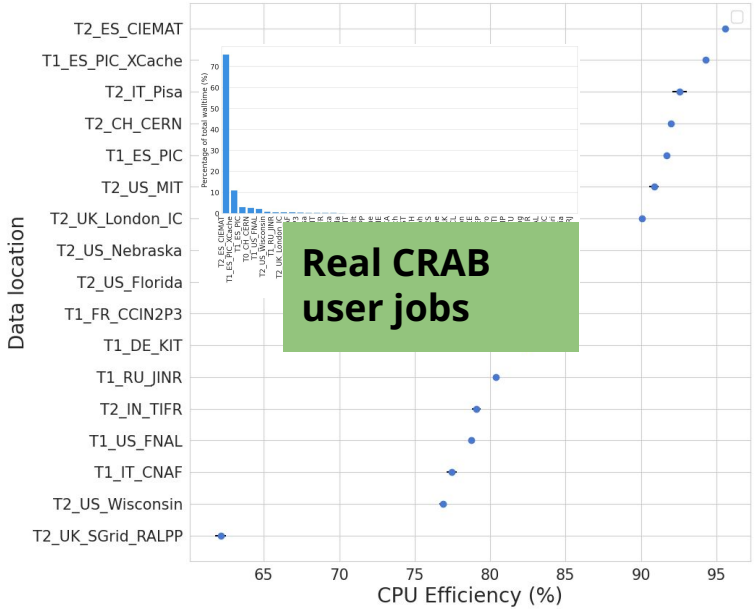
XCache **potentially reduce the amount of HS06·hr** spent by these tasks by **getting data closer to worker nodes being read as local** (latency hiding).

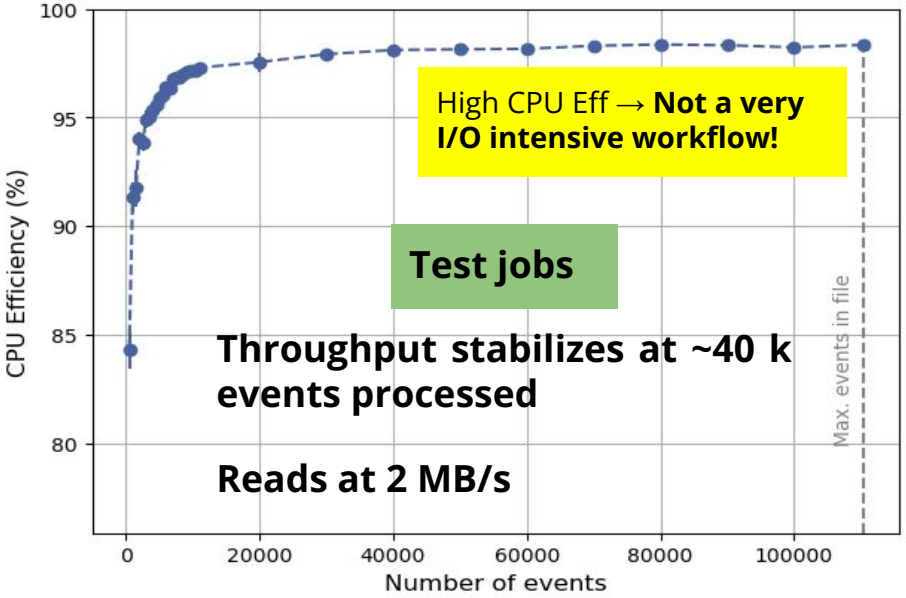
If the **access to XCache** was **enabled to the whole CIEMAT WNs,** the **13% of the total HS06·hr** spent **in CRAB jobs** that read data off the spanish region **would be reduced** → potential margin of improvement.



Studies on CPU efficiency dependence on network latency

In addition to analyzing real user jobs, **test jobs are submitted from PIC in a controlled environment**. These **test jobs access MINIAOD files from various CMS sites** distributed around the world.
→ MINIAOD files are compact versions of the full CMS event data used for analysis.



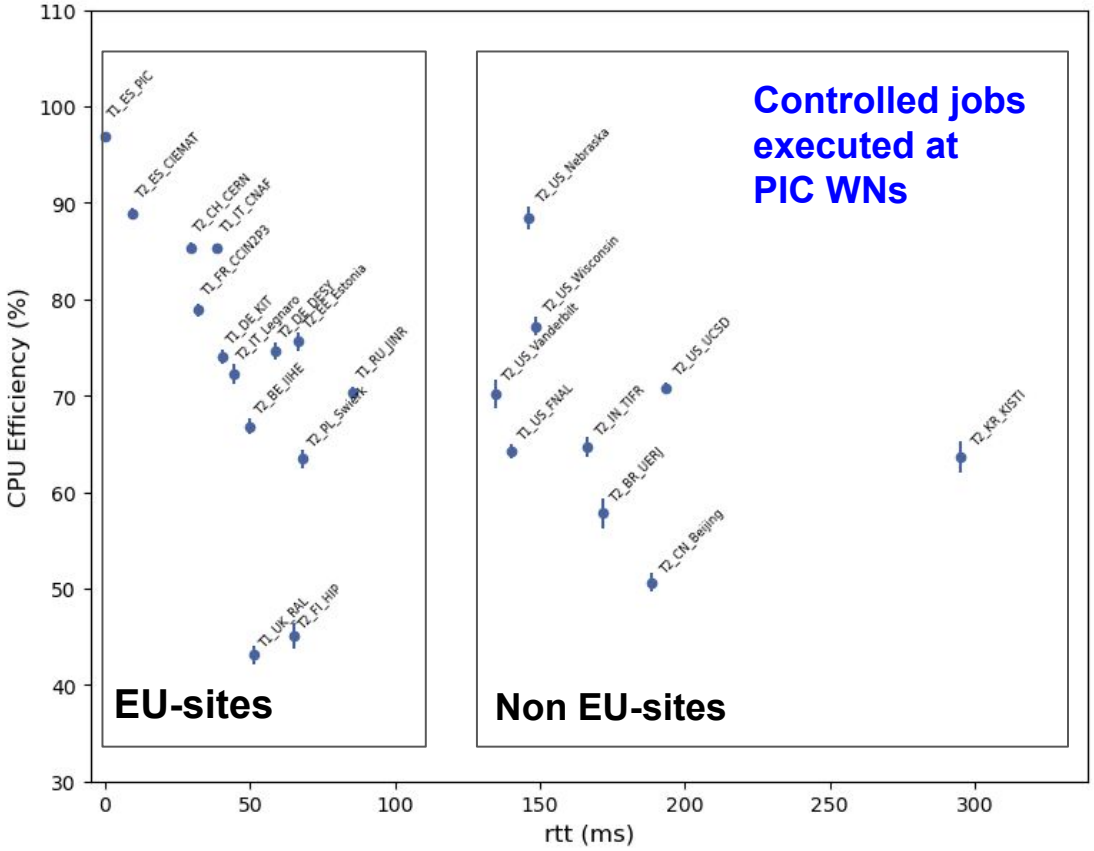


Executing **controlled jobs locally at PIC** allows us to **minimize the errors** and **study deeper aspects** of the jobs (ev. Processing throughput dependency on latency, i.e..).

Studies of CPU efficiency dependence on network latency

The goal of this study is to evaluate which is the **impact on CPU efficiency of executing the same job accessing similar MINIAOD files from different sites**.

Our measurement of CPU efficiency involved a **comparison** between executing the **same job** while reading **from XCache locally and remotely**, which reveals a **degradation on the CPU efficiency** → also, negative impact on walltime.

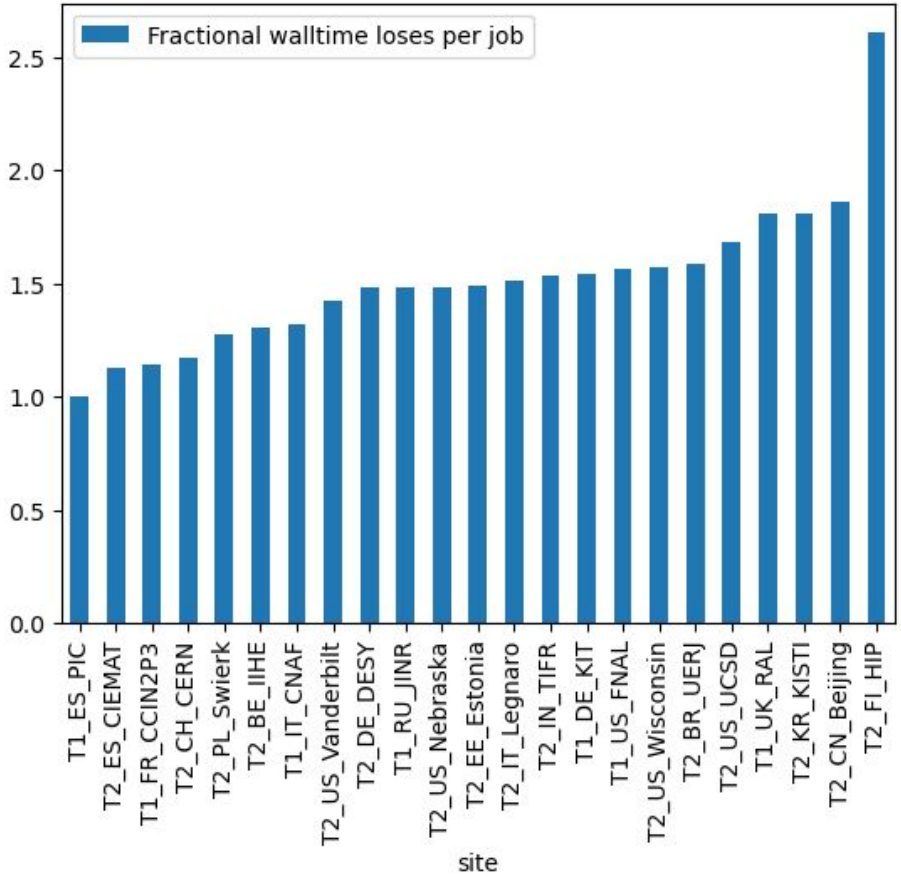


Walltime loss by jobs reading remote data

By comparing the time it takes to run jobs at the PIC versus other sites, **we calculate the wasted walltime using the number of jobs executed at each site.**

Overall, if all jobs were run accessing data locally at PIC, **1.8 kHS06·hours or 28% of walltime** could have been saved out of the **total 6.5k HS06·hours tested**.

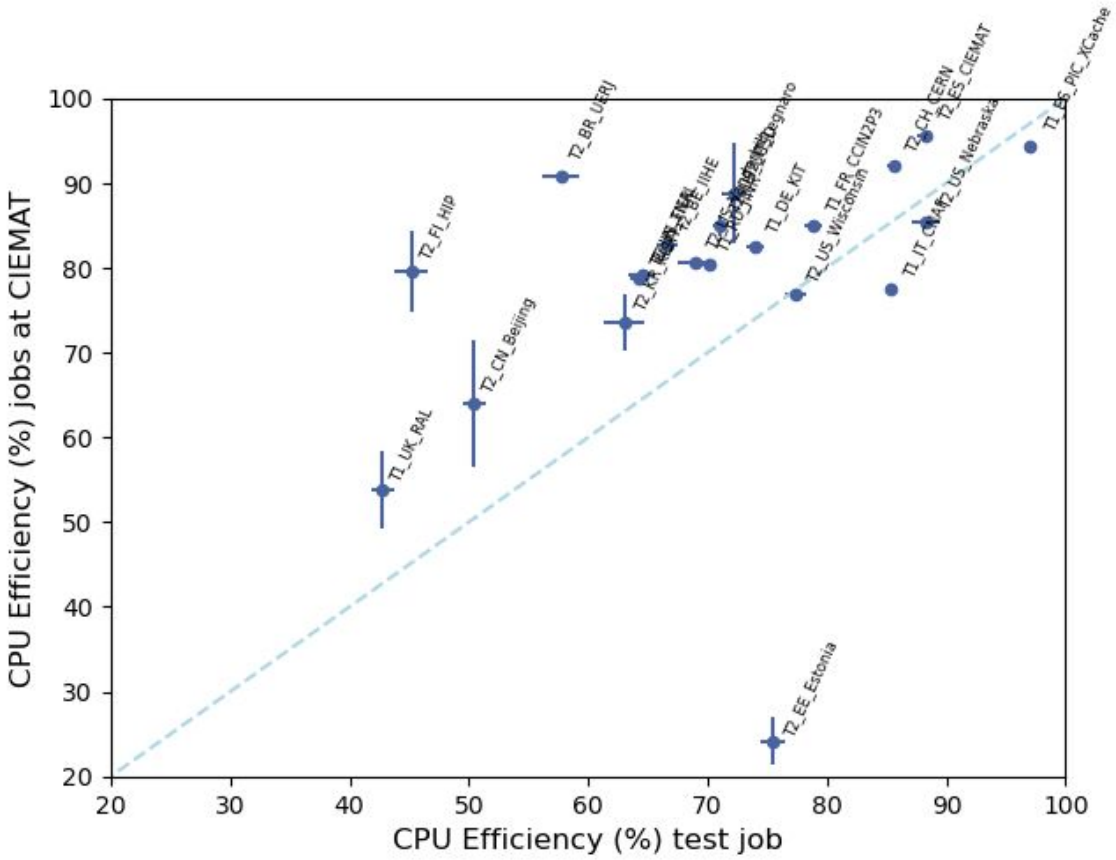
Total test: 6495 HS06·hour
Total wastage: 1812 HS06·hour
Fractional waste of walltime.: 28 %



Correlated CPU eff degradation in test and real jobs

This plot illustrates **the correlation between the degradation in CPU efficiency during remote reads for real user jobs (y-axis) and test jobs (x-axis)**

The relationship suggests a **relative tradeoff**, **verifying** that both the **degradation effect of CPU efficiency and the improvement in XCache are observable in reality** → validating our tests.

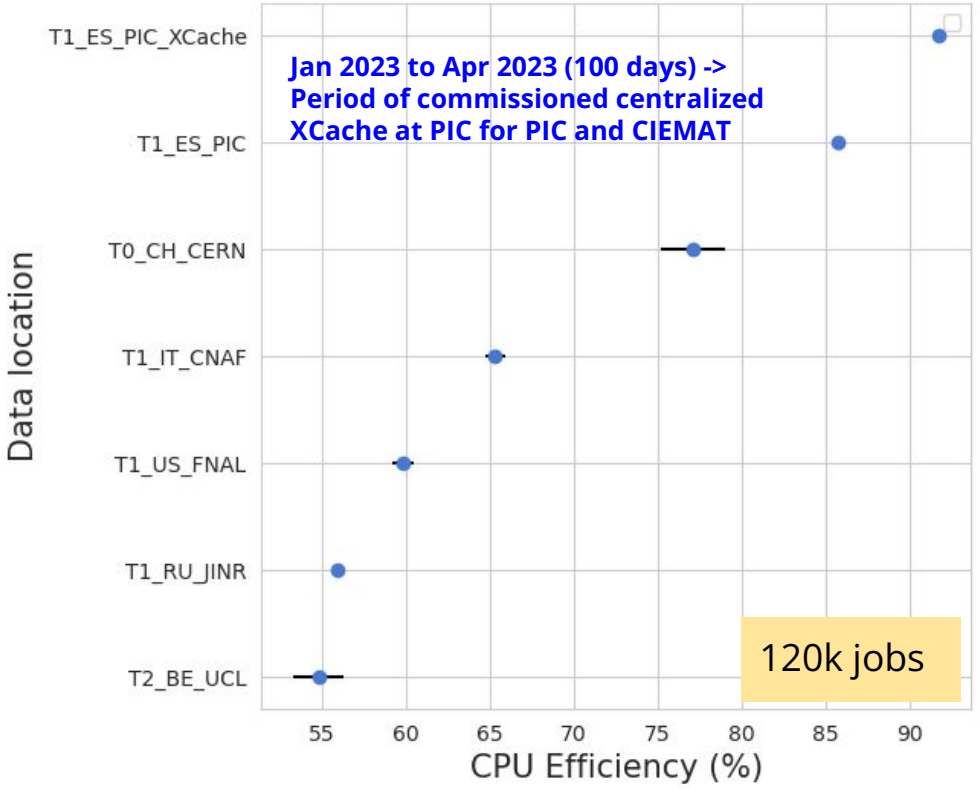


CRAB jobs executed at PIC worker nodes

Some users explicitly specify in the job from where they want the jobs to be read.

This procedure **bypasses the rules of the local site configuration**, which **hierarchically specifies to the jobs to download the data to the cache instead of reading them remotely** from any other site using the XRootd redirectors.

This is around 4k HS06·hr during the period **(1% of the total walltime in CRAB jobs)** → it is **less efficient and this data can not be cached.**



Conclusions

· XCache service has been deployed for PIC and CIEMAT CMS Tier sites to improve the overall CPU efficiency by reducing the latency to access remote input data.

· A monitoring service has been implemented to monitor the performance of the cache service in PIC, ensuring efficient and effective data delivery to the Spanish region.

· A significant improvement has been observed in the controlled jobs of analysis by reading MINIAOD data locally through XCache, compared to reading them remotely, as expected.

· In the period considered in this study, the 76% of walltime spent by CMS CRAB analysis jobs at CIEMAT read data locally. The XCache service provides significant benefits to the remaining 24% of walltime that reads data remotely, resulting in improved efficiency and saved walltime.

Outlook

· Future studies dedicated to XCache service are required to determine the extent of storage cost reduction achieved by retaining the most commonly accessed data for Analysis jobs within the cache.

· Centralized XCache for the region at PIC would help to alleviate the storage costs in the region, but network use will increase by sending data to CIEMAT. The impact over the network will also be evaluated in the future.

PIC
port d'informació científica

Thanks!
Questions?

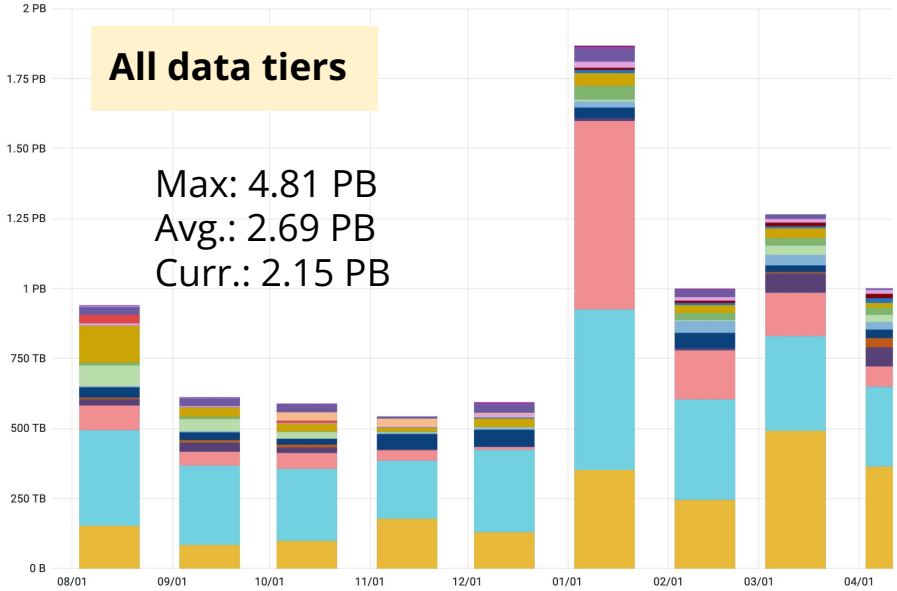
Projects FPA2016-80994-C2-1-R,
PID2019-110942RB-C21, and
BES-2017-082665 funded by:

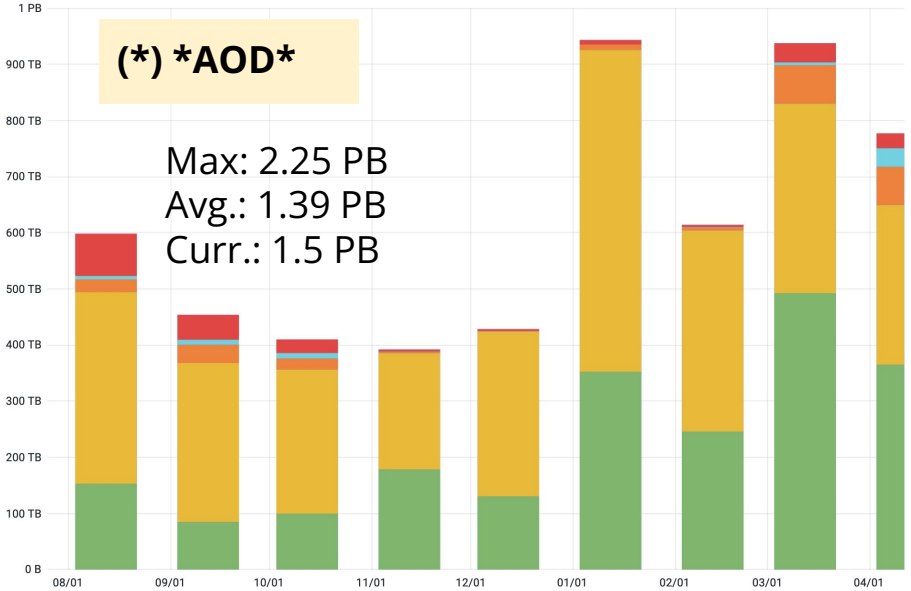
It has also been supported by the Ministerio de Ciencia e Innovación MCIN AEI/10.13039/501100011033 under contract PID2020-113614RB-C21, the Catalan government under contract 2021 SGR 00574, and the Red Española de Supercomputación (RES) through the grant DATA-2020-1-0039

BACKUPS

Data studies: Rucio usage 2022-2023 at PIC (3)

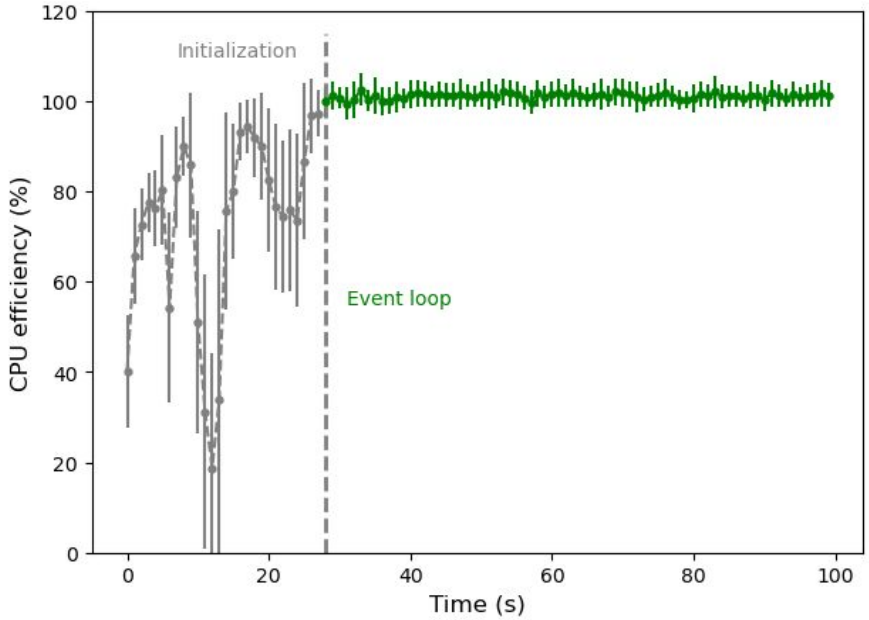
About half of the deployed disk storage is used for *AOD* derived formats

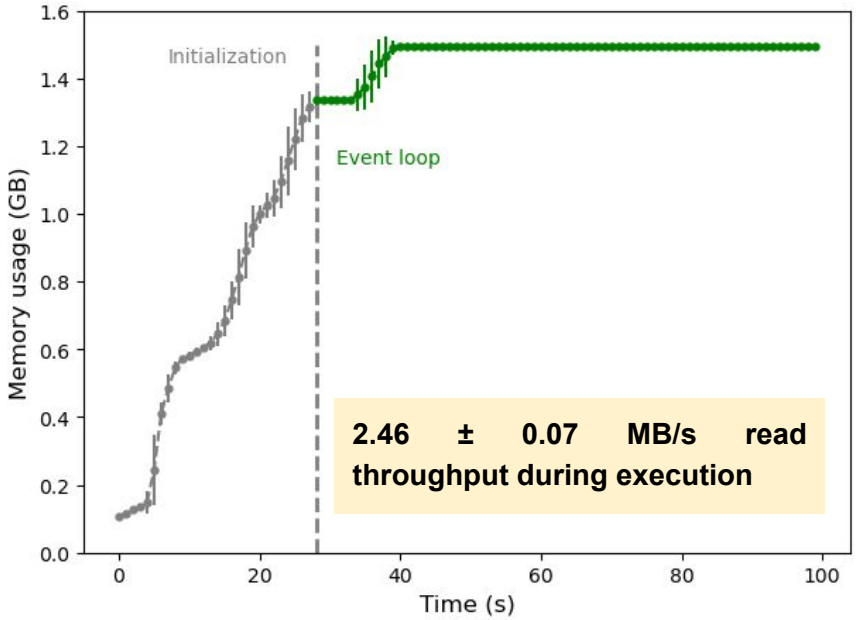




(*) *AOD* stands for all data tier including the word 'AOD' on it.

Controlled test job profiling executed at PIC





2.46 ± 0.07 MB/s read throughput during execution

Studies CPU efficiency dependance over latency