PIC
port d'informació
científica

CMS

Experience deploying xCache for CMS in Spain

Carlos Perez Dengra (CIEMAT-PIC), Josep Flix Molina (CIEMAT-PIC) and Anna Sikora (UAB) on behalf of the CMS Collaboration.

30 of March of 2023 (XRootd+FTS Workshop) @ Ljubljana

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** have the **capability to read data remotely** using the **CMS XRootD federation** (overflow to close sites, files opened in fallback and so on).
- We have been **exploring the xCache** service at PIC Tier-1 and CIEMAT Tier-2 centers to cache data which is read from remote centers
- xCache helps **reducing data access latency**, **improving CPU efficiency** and potentially **reducing the storage** deployed in the region
- We have deployed xCache services at both sites, and dedicated **studies and performance measurements** have been performed to configure the service

Commissioning the xCache (initial deployment)

- **Initially, two XCache services** were deployed at both PIC and CIEMAT to understand the service optimal configuration.

Deployment	Storage	XrootD version (OSG)	CMS data tiers to cache	Accessibility
Initial (2022-Feb 2023)	T1_ES_PIC: 150 TB (mult.disk) T2_ES_CIEMAT: 22 TB (RAID6)	5.4.2 5.4.2	**All** **All**	All compute nodes All compute nodes

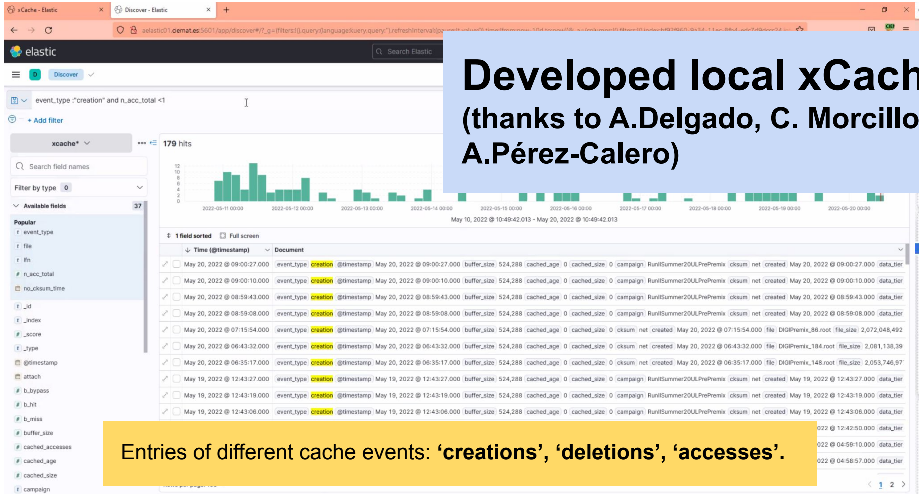
- **Caching all CMS data tiers Trivial File Catalog** configured to cache all data. **Inefficient**, but useful to **test the service at scale** → **not deployed this way anymore.**

- CMS pile-up data produced **high loads in the cache and memory misuse that resulted in high jobs failure rates** → **reported to developers**

- **Local monitoring** system deployed based, parsing .cinfo files (next slide)

Developed local xCache monitoring service

(thanks to A.Delgado, C. Morcillo Perez, Fco. Rodríguez and A.Pérez-Calero)



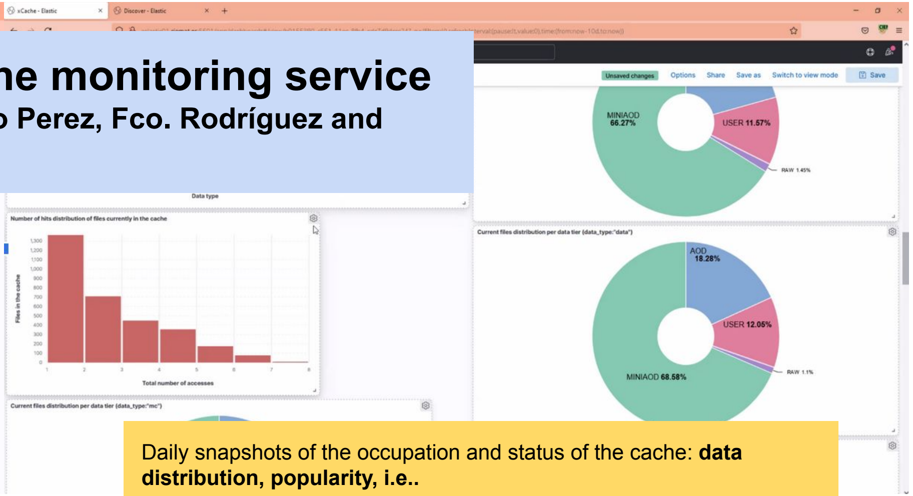
event_type :"creation" and n_acc_total <1

Fields: event_type, file, lfn, n_acc_total, no_cksum_time, _id, _index, _score, _type, @timestamp, attach, b_bypass, b_hit, b_miss, buffer_size, cached_accesses, cached_age, cached_size, campaign

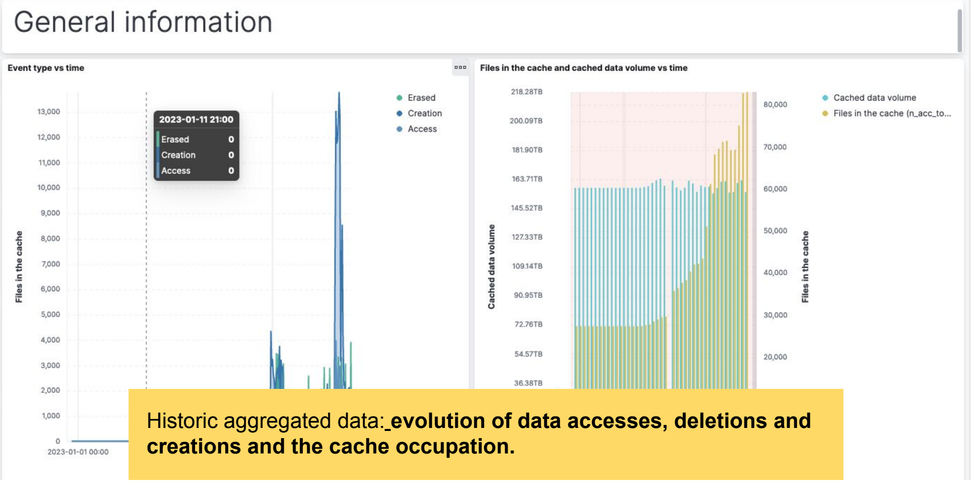
May 10, 2022 @ 10:49:42.013 - May 20, 2022 @ 10:49:42.013

```
Time (@timestamp)           Document
May 20, 2022 @ 09:00:27.000 event_type creation @timestamp May 20, 2022 @ 09:00:27.000 buffer_size 524,288 cached_age 0 cached_size 0 campaign RunIISummer20ULPrePremix cksum net created May 20, 2022 @ 09:00:27.000 data_tier
May 20, 2022 @ 09:00:10.000 event_type creation @timestamp May 20, 2022 @ 09:00:10.000 buffer_size 524,288 cached_age 0 cached_size 0 campaign RunIISummer20ULPrePremix cksum net created May 20, 2022 @ 09:00:10.000 data_tier
May 20, 2022 @ 08:59:43.000 event_type creation @timestamp May 20, 2022 @ 08:59:43.000 buffer_size 524,288 cached_age 0 cached_size 0 campaign RunIISummer20ULPrePremix cksum net created May 20, 2022 @ 08:59:43.000 data_tier
May 20, 2022 @ 08:59:08.000 event_type creation @timestamp May 20, 2022 @ 08:59:08.000 buffer_size 524,288 cached_age 0 cached_size 0 campaign RunIISummer20ULPrePremix cksum net created May 20, 2022 @ 08:59:08.000 data_tier
May 20, 2022 @ 07:15:54.000 event_type creation @timestamp May 20, 2022 @ 07:15:54.000 buffer_size 524,288 cached_age 0 cached_size 0 cksum net created May 20, 2022 @ 07:15:54.000 file DIGIPremix_86.root file_size 2,072,048,492
May 20, 2022 @ 06:43:32.000 event_type creation @timestamp May 20, 2022 @ 06:43:32.000 buffer_size 524,288 cached_age 0 cached_size 0 cksum net created May 20, 2022 @ 06:43:32.000 file DIGIPremix_184.root file_size 2,081,138,39
May 20, 2022 @ 06:35:17.000 event_type creation @timestamp May 20, 2022 @ 06:35:17.000 buffer_size 524,288 cached_age 0 cached_size 0 cksum net created May 20, 2022 @ 06:35:17.000 file DIGIPremix_148.root file_size 2,053,746,97
May 19, 2022 @ 12:43:27.000 event_type creation @timestamp May 19, 2022 @ 12:43:27.000 buffer_size 524,288 cached_age 0 cached_size 0 campaign RunIISummer20ULPrePremix cksum net created May 19, 2022 @ 12:43:27.000 data_tier
May 19, 2022 @ 12:43:19.000 event_type creation @timestamp May 19, 2022 @ 12:43:19.000 buffer_size 524,288 cached_age 0 cached_size 0 campaign RunIISummer20ULPrePremix cksum net created May 19, 2022 @ 12:43:19.000 data_tier
May 19, 2022 @ 12:43:06.000 event_type creation @timestamp May 19, 2022 @ 12:43:06.000 buffer_size 524,288 cached_age 0 cached_size 0 campaign RunIISummer20ULPrePremix cksum net created May 19, 2022 @ 12:43:06.000 data_tier
022 @ 12:42:50.000 data_tier
022 @ 04:59:10.000 data_tier
022 @ 04:58:57.000 data_tier
```

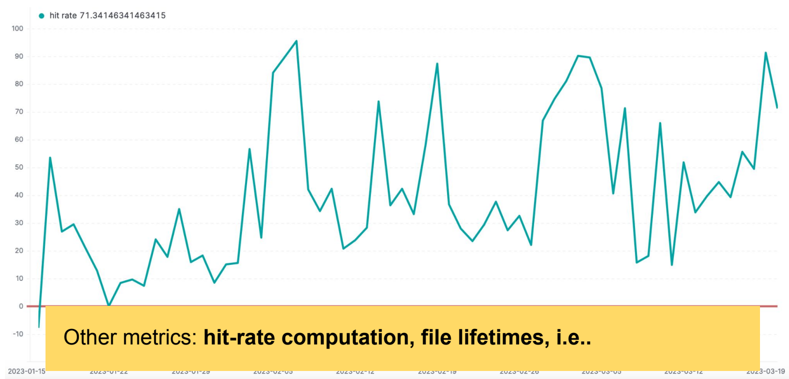
Entries of different cache events: **‘creations’, ‘deletions’, ‘accesses’.**



Daily snapshots of the occupation and status of the cache: **data distribution, popularity, i.e..**



Historic aggregated data: **evolution of data accesses, deletions and creations and the cache occupation.**

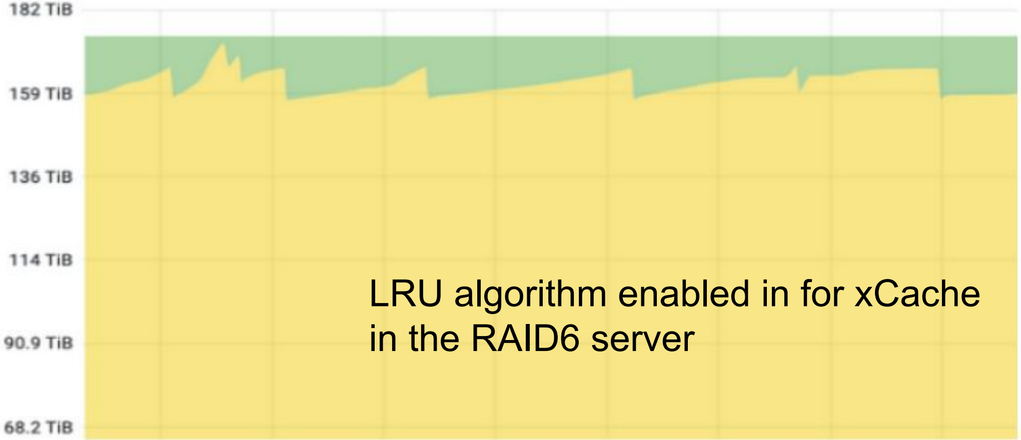


Other metrics: **hit-rate computation, file lifetimes, i.e..**

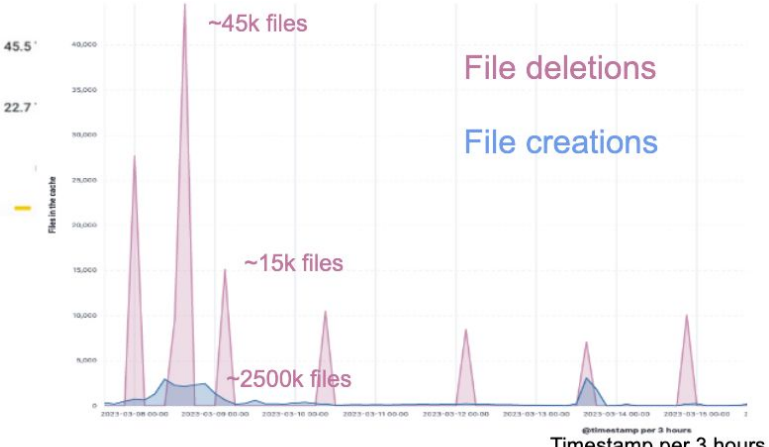
XCache deployed in PIC (current deployment)

Currently, a **single cache at PIC (180TB)** is serving data for both **PIC and CIEMAT** compute nodes → executed tasks from both sites WN's **reading remote input data are effectively running.**

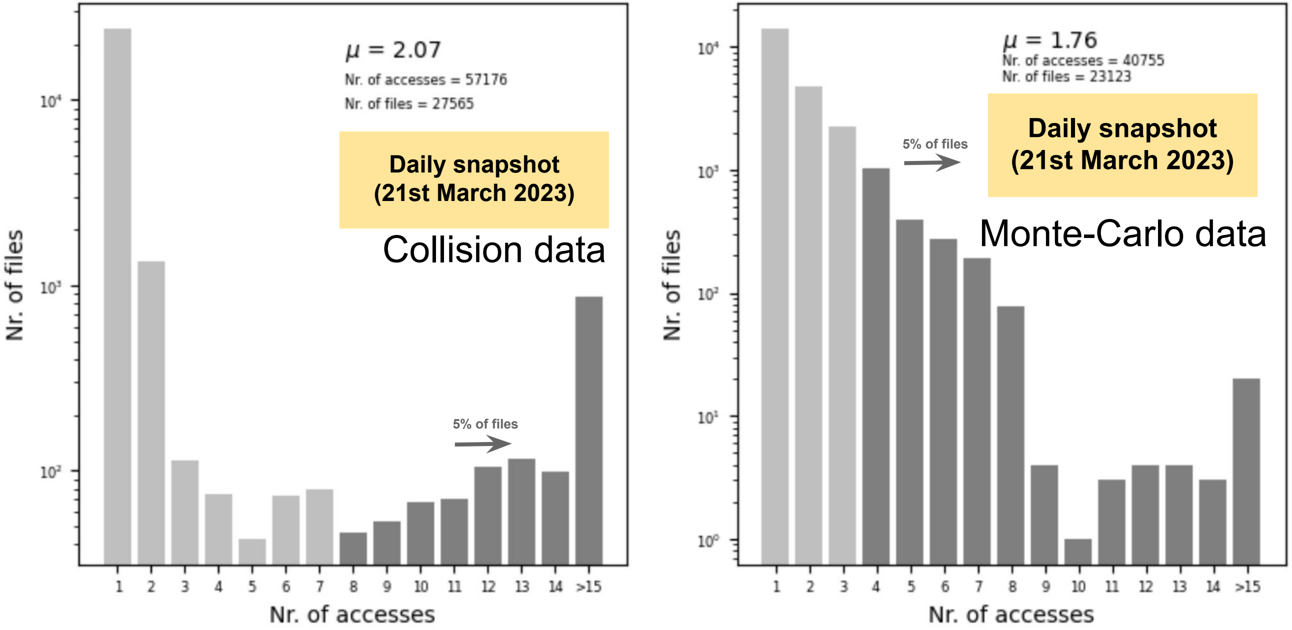
Deployment	Storage	XrootD version (OSG)	CMS data tiers to cache	Accessibility
Current (single cache for both sites)	T1_ES_PIC: 180 TB (RAID)	5.5.1	**All (except PREMIX, "*PrePremix*" and unmerged)**	All compute nodes + half of the compute nodes at CIEMAT



LRU algorithm enabled in for xCache in the RAID6 server

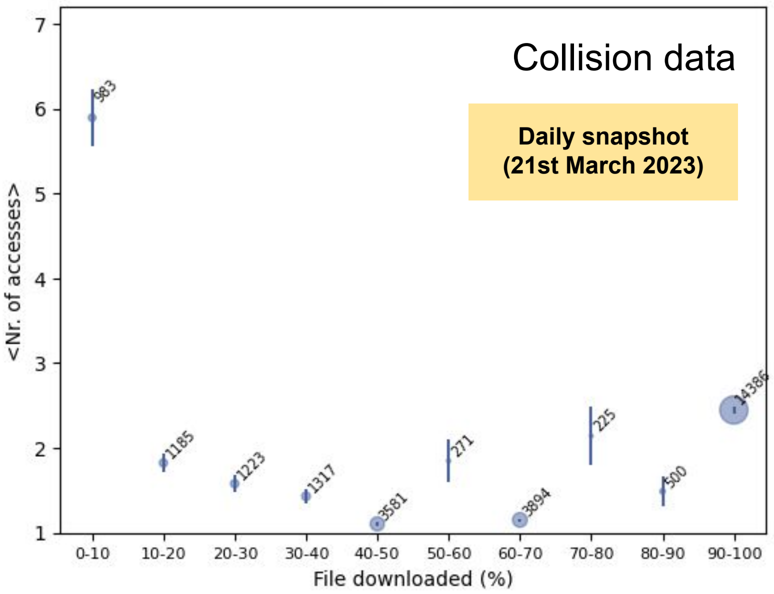


Data popularity at PIC XCache

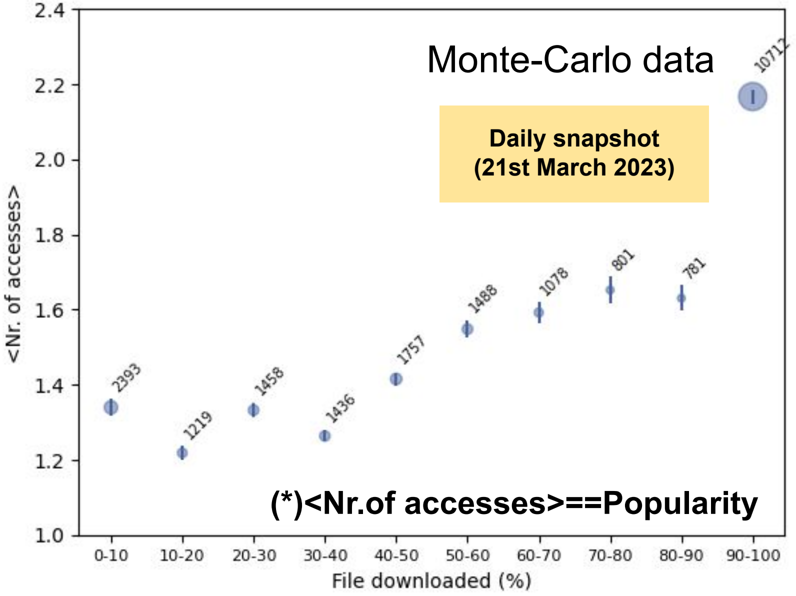


Average number of files accesses for both 'data' and 'mc' file types stored in the cache. This corresponds to a daily snapshot taken from .cinfo files

Files download percentage at PIC XCache


Collision data

Daily snapshot (21st March 2023)


Monte-Carlo data

Daily snapshot (21st March 2023)

(*)<Nr.of accesses>==Popularity

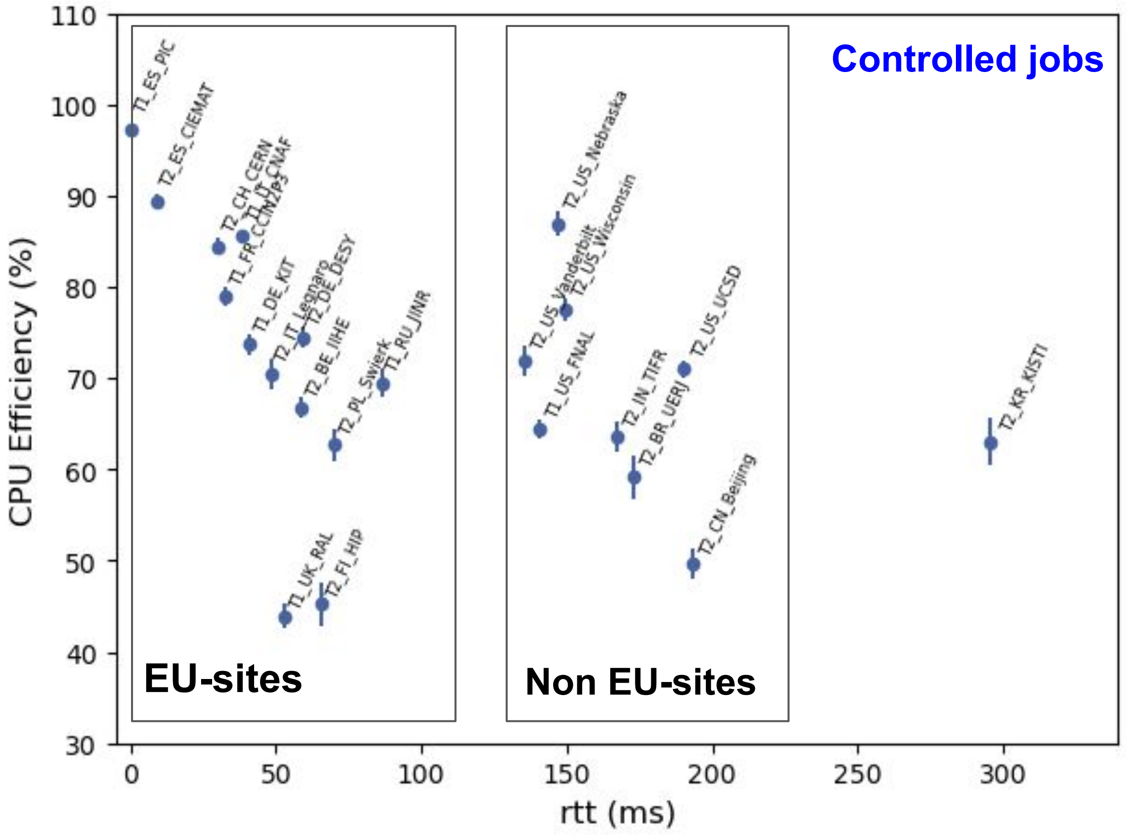
Collision and Monte-Carlo data show a **different behavior of re-accessibility.** In both cases**, the intermediate partially downloaded files are less re-accessed.**

CPU efficiency dependence of remote reads

We conducted controlled tests to assess the qualitative **dependence of CPU efficiency on remote reads** from **various CMS sites** in and outside of Europe

- Our test analysis jobs **read MiniAOD files from several sites** world wide. Note, however, that **this is not how CMS generally undertakes analysis workflows** → CMS late-binding.

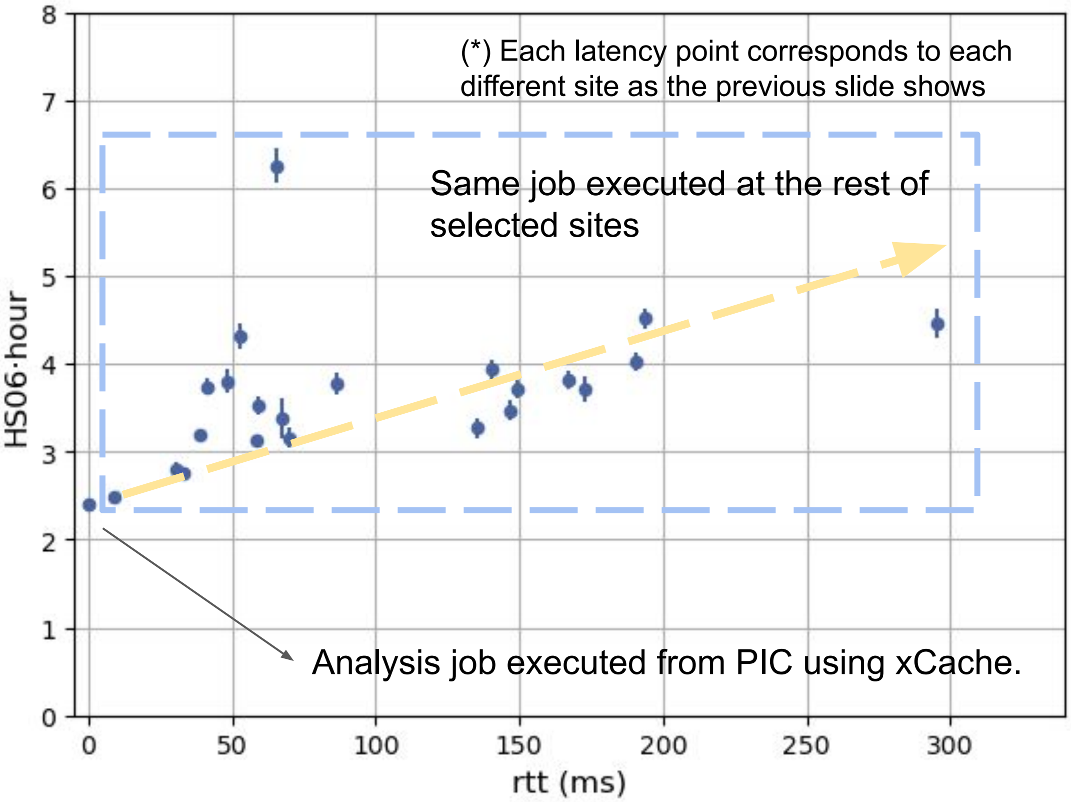
- **Controlled, uniform test** jobs allow us to **minimize errors** and make comparisons between source sites.



CPU efficiency dependence of remote reads (2)

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

We have measured how **the CPU efficiency improves by executing** the same job reading **locally from the xCache** compared to doing it remotely -> **HS06·hours ‘loss’** during the execution (see the figure)



(*) Each latency point corresponds to each different site as the previous slide shows

Same job executed at the rest of selected sites

Analysis job executed from PIC using xCache.

Conclusions

-During the initial phase of our project, we deployed the XCache from both PIC and CIEMAT separately to compare and validate the effects of different configurations on the service.

-Subsequently, we deployed the monitoring service and unified the service in PIC to efficiently and effectively serve data to the Spanish region.

- There is still some room to improve the re-accessibility of cached data →
Update the the site configuration with the actual data popularity of the service in production.

-The results of executing controlled analysis jobs remotely accessing MINIAOD at various sites reveal a significant degradation of CPU efficiency with latency due network and distance (as expected)

Outlook

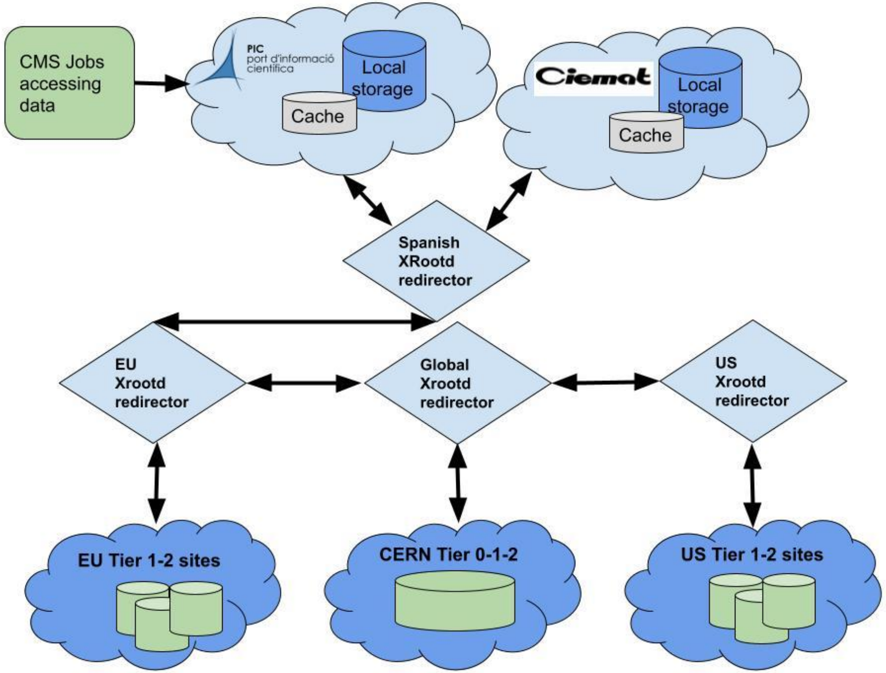
-The ongoing studies over the degradation of CPU efficiency will be complemented by dedicated studies evaluating the potential walltime saved by caching these data.

-Also, bringing the data closer to the cache increase the CPU efficiency of those jobs on our nodes. This effect has also to be well evaluated in jobs in production.

-Since xCache service is expected to alleviate the storage costs in the future, more dedicated studies will tell us how much storage we can save by keeping the most popular data for Analysis jobs within the cache.

Backup slides

XRootd redirectors infrastructure with the inclusion of the XCache service in PIC and CIEMAT sites

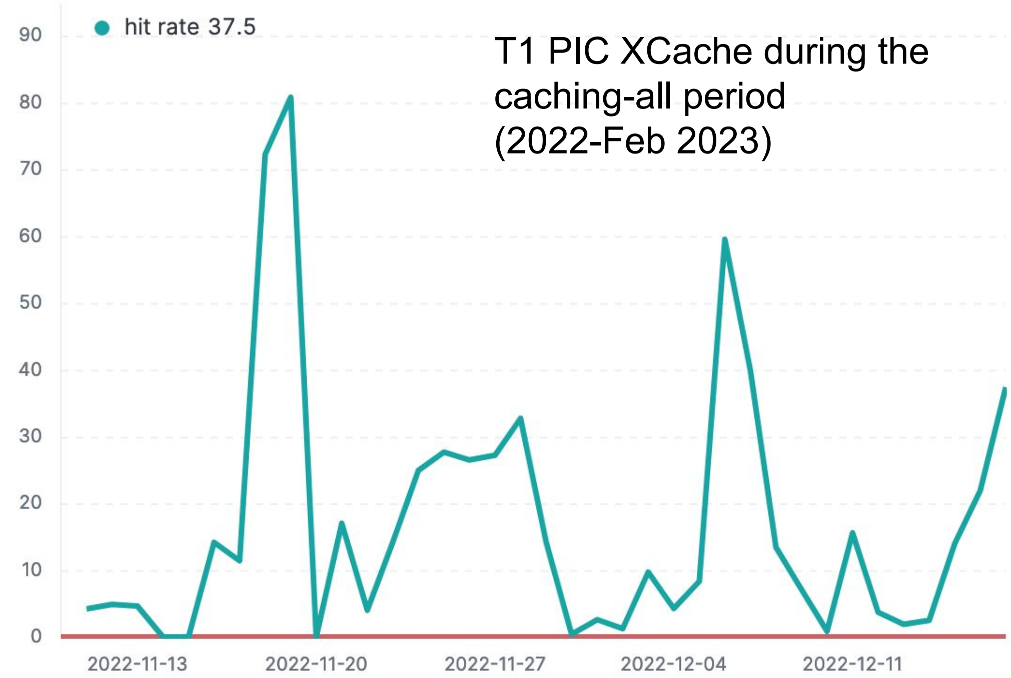


Conclusions and actions taken in the first period (2022-Feb 2023)

Due the incidents experienced, **we decided to exclude PREMIX from the cache:** CMS jobs access the PREMIX data in non-uniform way / only used once or very infrequently → **(see the Hit-rate for PREMIX period at PIC figure).**

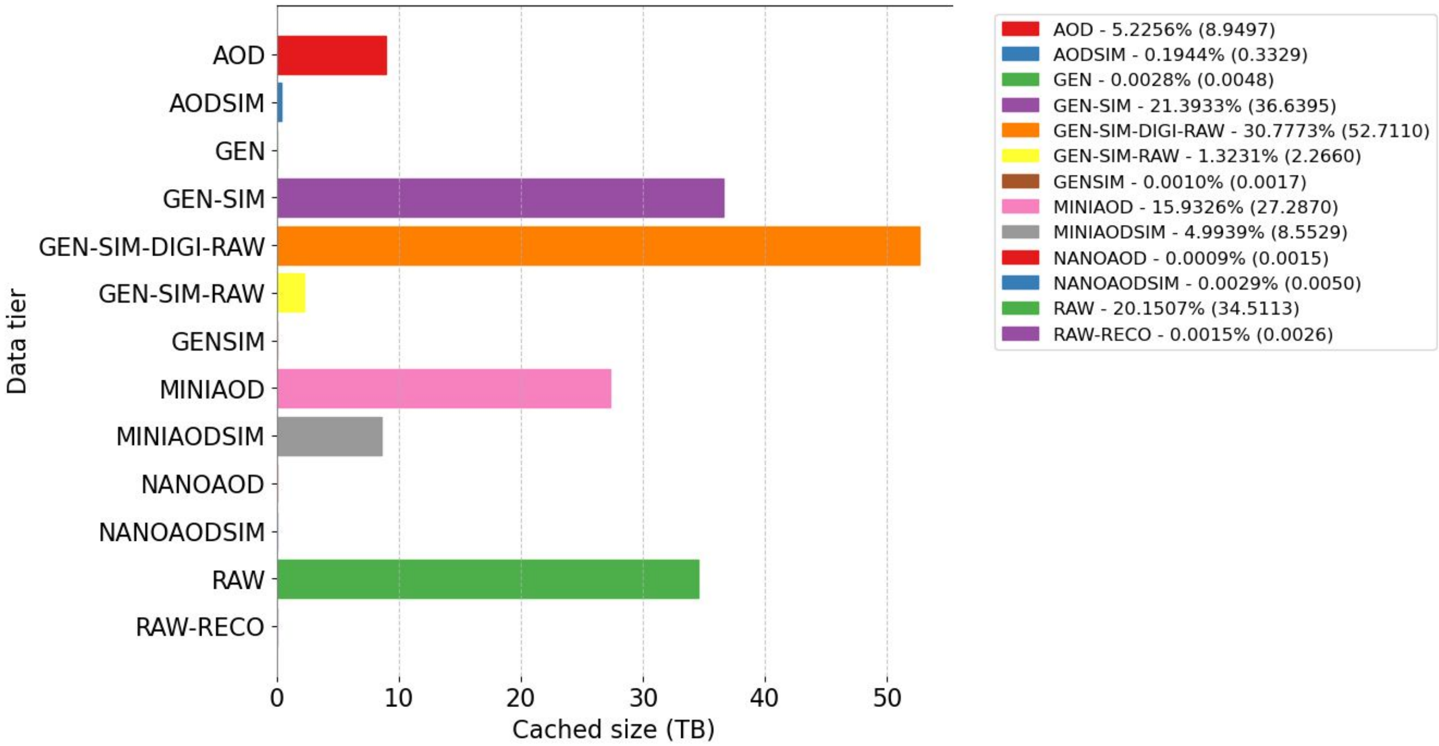
Similar approach was made for **PrePremix and unmerged data**, since several Processing jobs also failed due accessing these data.

Hit rate vs time

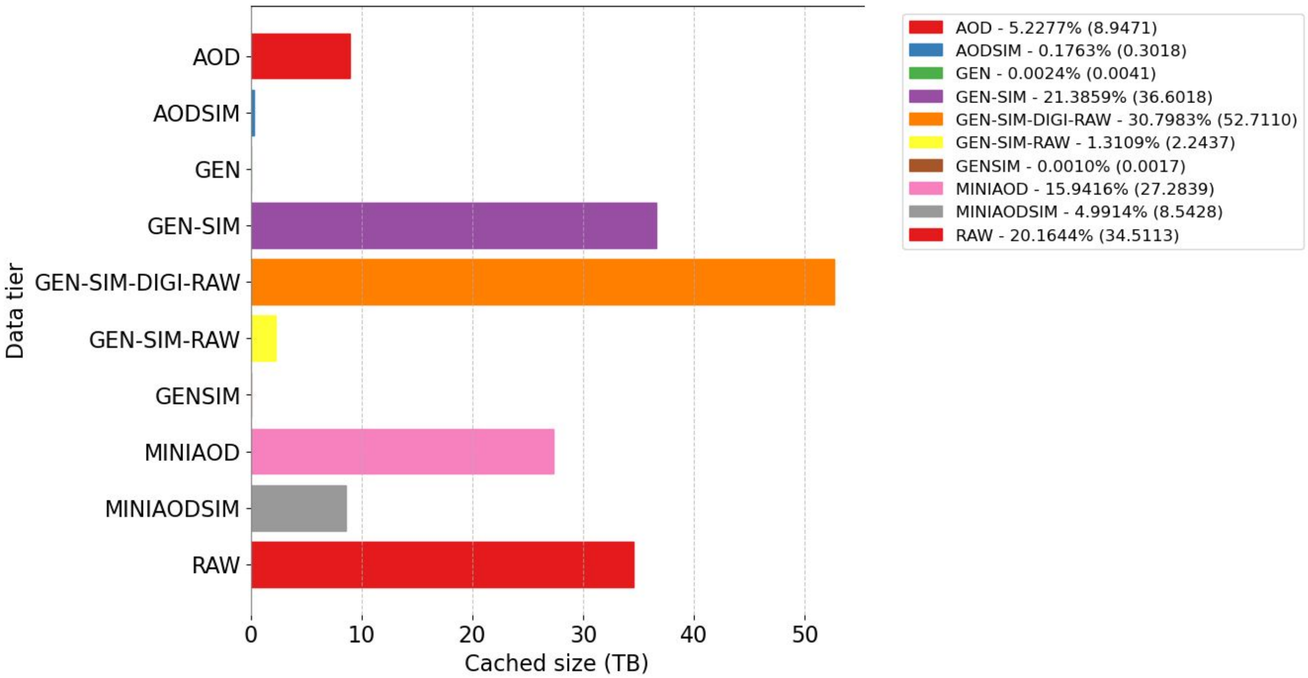


T1 PIC XCache during the caching-all period (2022-Feb 2023)

Distribution of data cached by XCache for PIC+CIEMAT

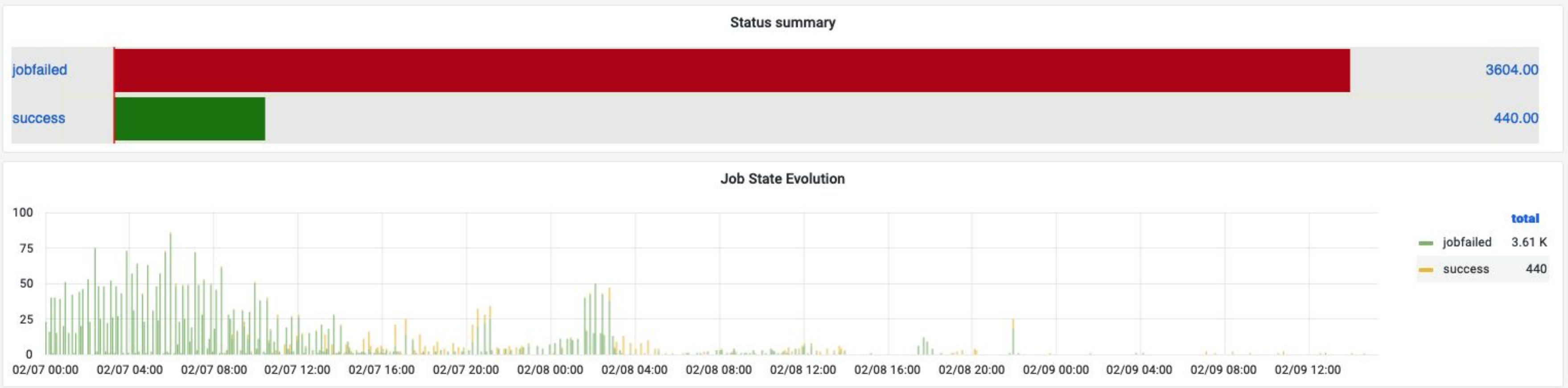


Distribution of data cached by XCache for PIC+CIEMAT without unmerged

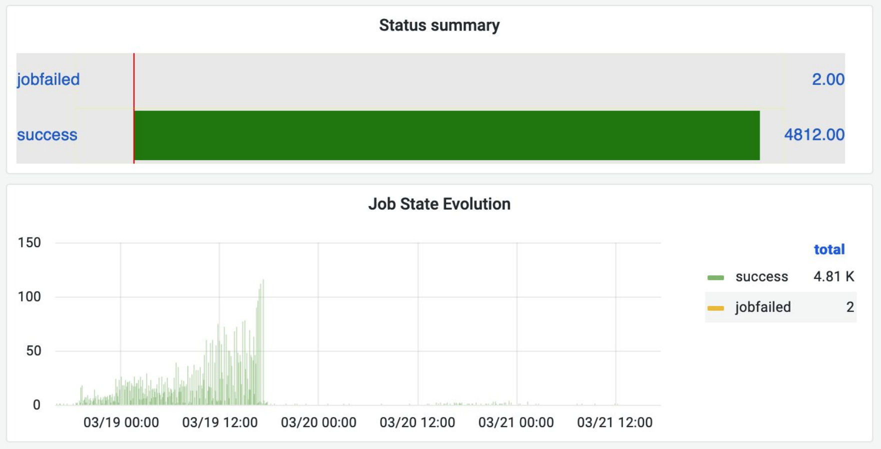


High failure rate Processing jobs

Caching PrePremix and unmerged

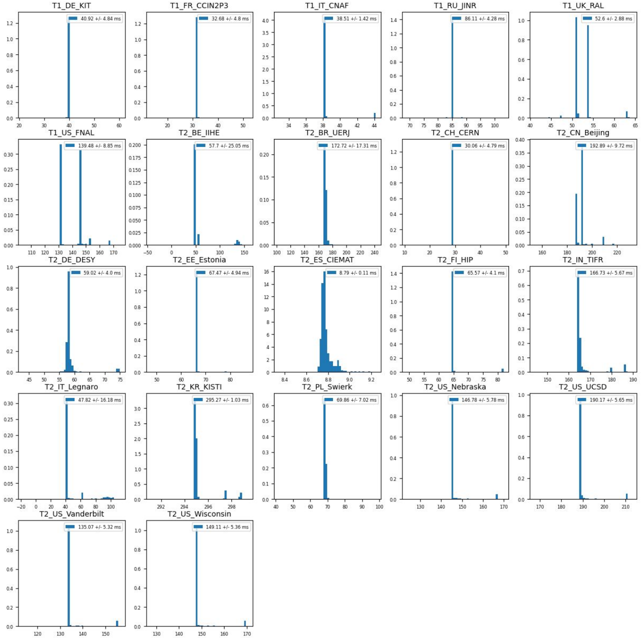


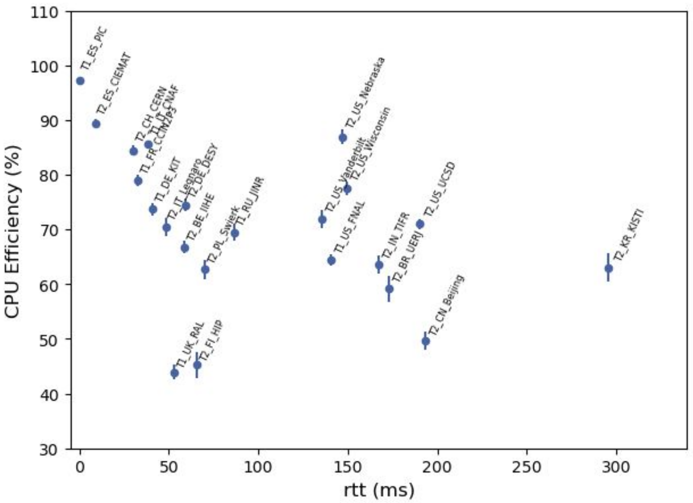
Currently excluding PrePremix and unmerged ->

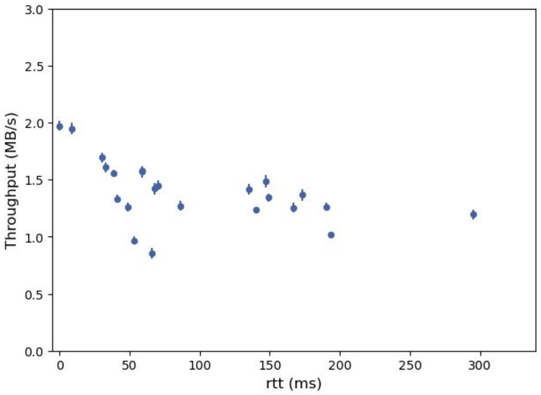


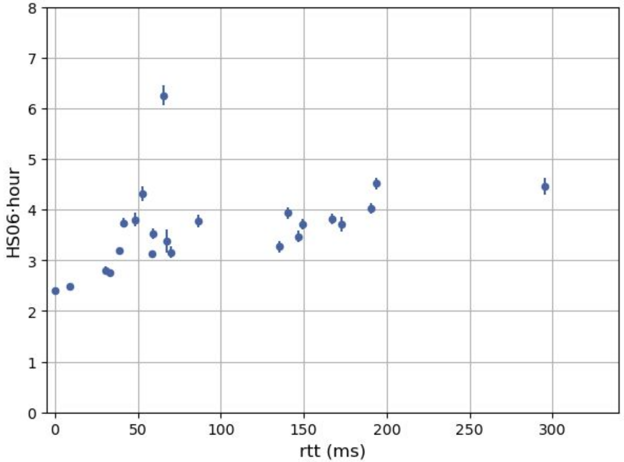
Plots (PEPE)

Remote sites (3200 HS06·hours tests)





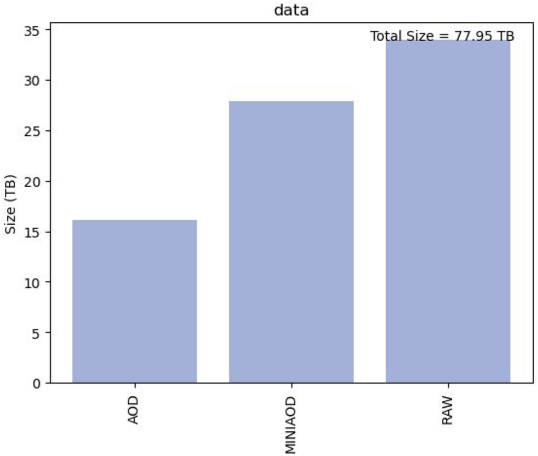


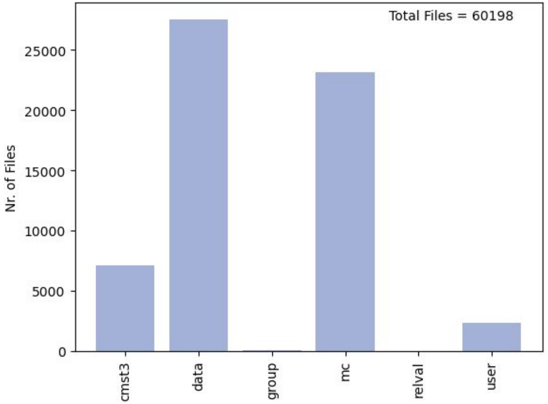


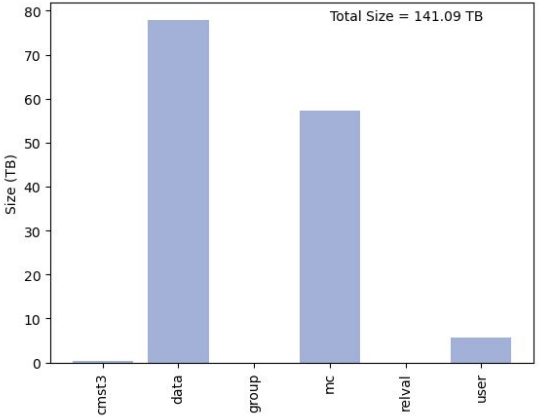
Cache content

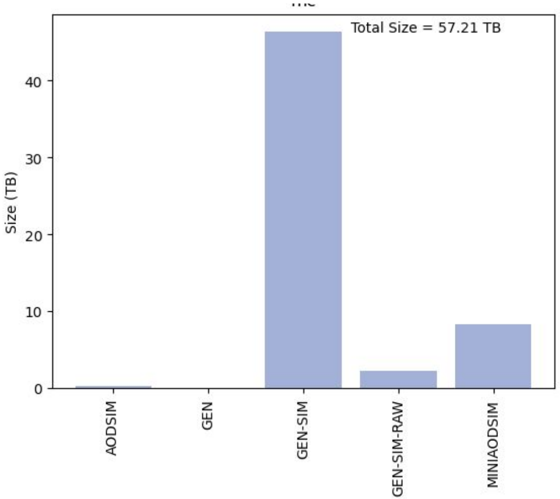
He filtrado

- fichero de tests que usamos en los tests controlados de Muon Analysis
- /store/test, /store/unmerged, *PrePremix*
- Todos aquellos ficheros en los que no se ha bajado ningún byte

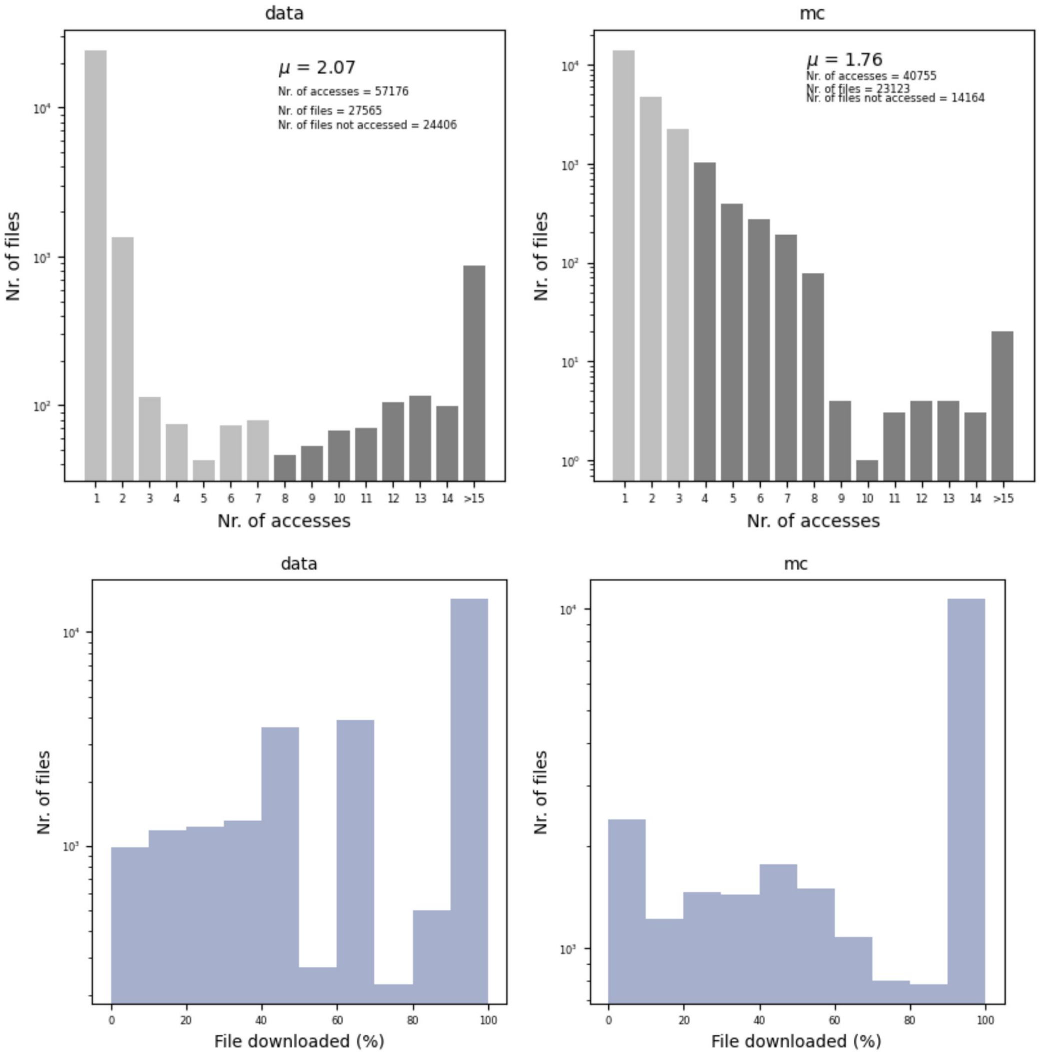




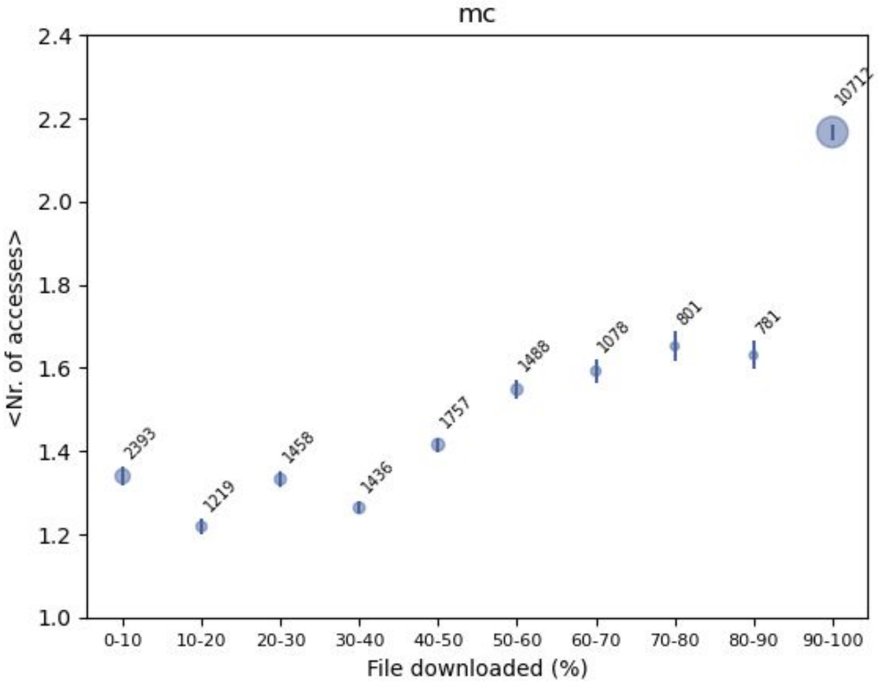


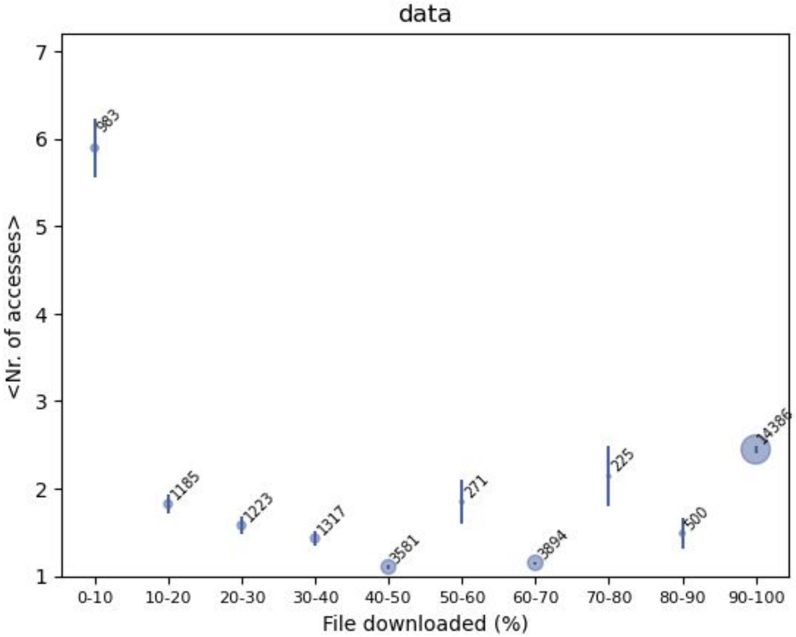


Cache content

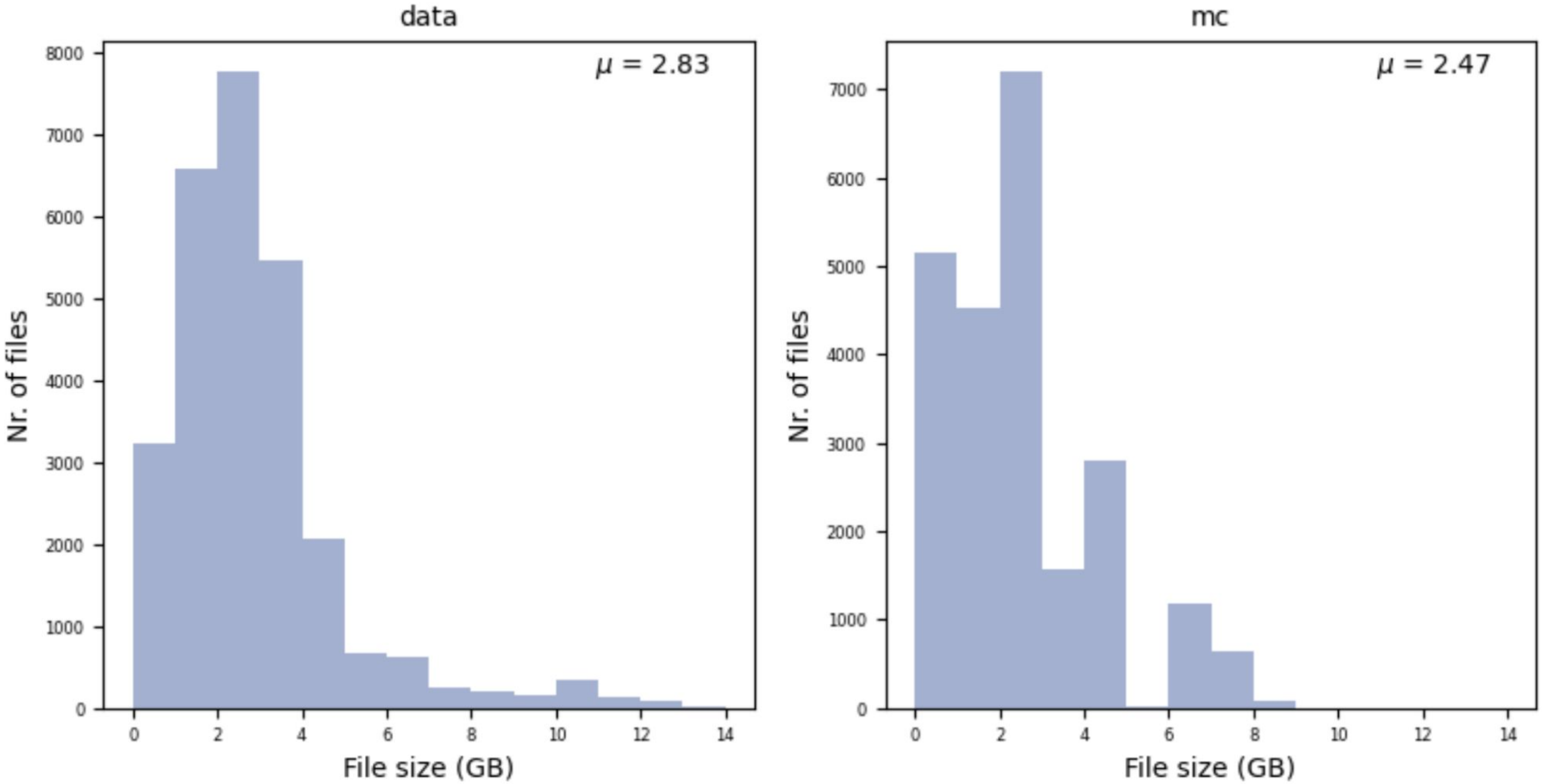


Cache content

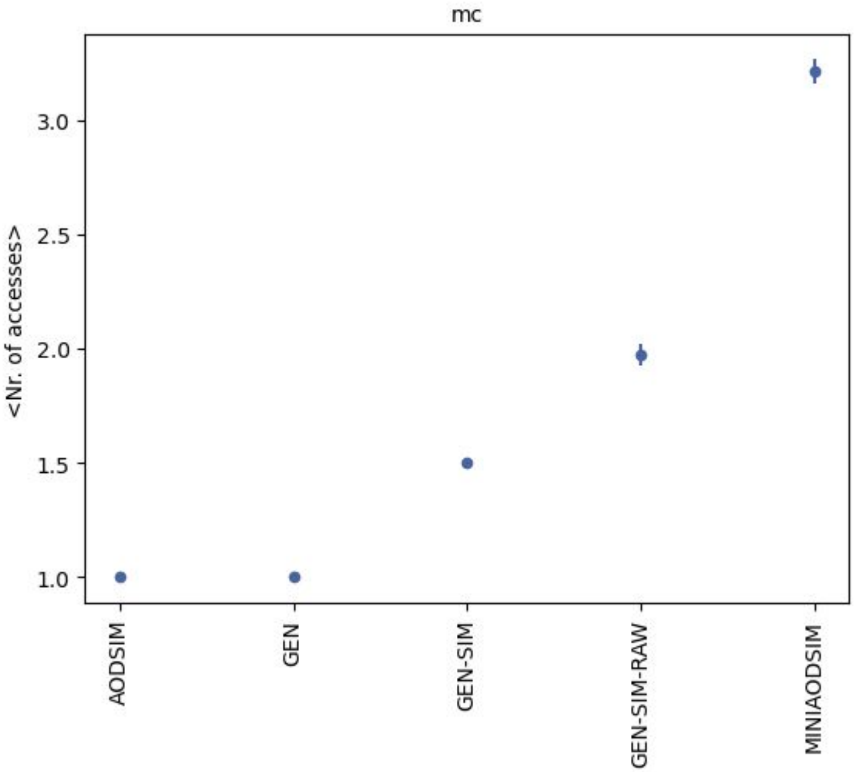


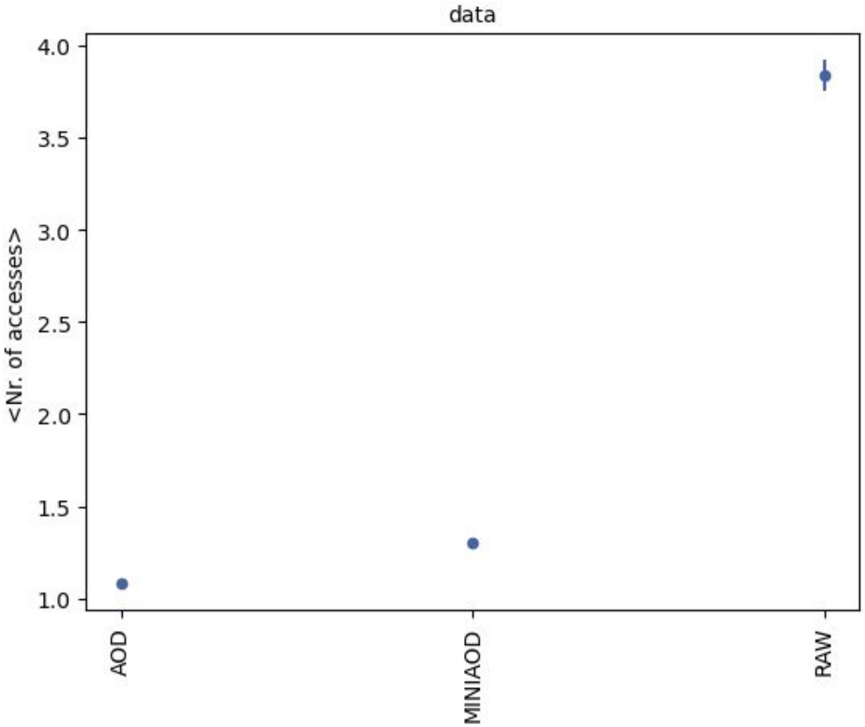


Cache content



Cache content


mc
<Nr. of accesses>
AODSIM
GEN
GEN-SIM
GEN-SIM-RAW
MINIAODSIM


data
<Nr. of accesses>
AOD
MINIAOD
RAW