LHCb Computing Resource usage in 2023

LHCb Public Note

Issue: First version
Revision: 0

Reference: LHCb-PUB-2024-003
Created: 7th January 2024
Last modified: 13th February 2024

Prepared By: LHCb Computing Project
C. Bozzi/Editor

Abstract

This document reports the usage of computing resources by the LHCb collaboration during the period January 1st – December 31st, 2023.

The data in the following sections have been compiled from the WLCG Accounting Utility (WAU) portal: https://monit-grafana.cern.ch/dashboards/f/Jux9_-1ik/accounting. For LHCb-specific information, the data is taken from the DIRAC Accounting at the LHCb DIRAC Web portal: https://lhcb-portal-dirac.cern.ch/.

Table of Contents

1. INTRODUCTION 1
2. COMPUTING ACTIVITIES DURING 2023 2
3. USAGE OF CPU RESOURCES 4
3.1. WLCG ACCOUNTING 5
3.2. LHCB DIRAC CPU ACCOUNTING 8
3.3. USAGE OF FAST SIMULATIONS AND FILTERED PRODUCTIONS 14
3.4. DIFFERENCES BETWEEN WLCG AND DIRAC ACCOUNTING 15
4. USAGE OF STORAGE RESOURCES 16
4.1. TAPE STORAGE 16
4.2. DISK AT TIER0, TIER1S AND TIER2S 17
5. DATA POPULARITY 23
6. SUMMARY 27
APPENDIX A 28

List of Figures

Figure 3-1: Summary of LHCb computing activities (used CPU power) from Jan 1st to December 31st, 2023. The line represents a weighted average of the 2022 and 2023 pledges. 4

Figure 3-2: Summary of LHCb site usage from Jan 1st until December 31st, 2023. 5

Figure 3-3: Average monthly CPU power provided by the Tier1s (and Tier0) to LHCb from Jan 1st until December 31st, 2023. 6

Figure 3-4: Average monthly CPU power provided by the Tier2s to LHCb from Jan 1st until October 31st, 2023. 7

Figure 3-5: CPU time consumed by LHCb from Jan 1st until December 31st, 2023 (in days), for WLCG sites, excluding the HLT farm (top, per country), and non-WLCG sites, including the HLT farm (bottom, per site). 9

Figure 3-6: Usage of LHCb Tier0/1s during Jan 1st – Dec 31st, 2023. The top plot shows the used CPU power as a function of the different activities, while the bottom plot shows the contributions from the different sites. User jobs and analysis productions (shown in magenta and light green in the top plot) are further detailed in Figure 3-8. 10

Figure 3-7: Usage of resources in WLCG sites other than Tier0 and Tier1s, during Jan 1st— Dec 31st, 2023. The top plot shows the used CPU power as a function of the different activities, while the bottom plot shows the contributions from the different sites. User jobs (shown in magenta in the top plot) are further detailed in Figure 3-8. 11

Figure 3-8: Running user and analysis production jobs on Tier0 and Tier1 sites (top left), and outside (top right) during Jan 1st – Dec 31st, 2023. The two lower plot show the contribution of the two workflows (user jobs in magenta, analysis productions in light green) 12

Figure 3-9: CPU time as a function of the job final status for all jobs (top); for failed jobs, the middle and bottom plots show the error status and activity, respectively. 13

Figure 3-10: Left: number of events produced in 2023 by simulation type; right: number of events produced per day, with rolling averages over 7, 30, 90 days. 14

Figure 3-11: Number of events simulated as a function of calendar year. Each color corresponds to the simulation of a given data-taking year. 14

Figure 4-1: Tape space occupancy for RAW (left) and RDST (right) data, January 1st – December 31st, 2023. The glitch at the end of January in this and in the following plots corresponds to an outage of the LHCb accounting system and it is not significant. 16

Figure 4-2: Tape occupancy for derived data, January 1st – December 31st, 2023. The glitch at the end of January corresponds to an outage of the LHCb accounting system and it is not significant. 16

Figure 4-3: Usage of Disk resources at CERN and Tier1s from Jan 1st to December 31st, 2023. Real data and simulated data are shown in the top and bottom figures. 19

Figure 4-4: Usage of LHCb buffer disk from January 1st to December 31st, 2023. 20

Figure 4-5: Usage of disk resources reserved for Users from January 1st to December 31st, 2023. 20

Figure 4-6: Usage of disk space at Tier-2D sites from January 1st to December 31st, 2023, for real data (top) and simulated data (bottom). 22

Figure 5-1: Volume (as percentage of total data volume) of datasets accessed at least once in the last 13, 26, 52 weeks, in various periods as reported in this and previous documents. 23

Figure 5-2: Volume of accessed datasets (in TB) as a function of the number of accesses in the last 13 weeks. Top: physical volume, bottom: logical volume 24

Figure 5-3: Volume of accessed datasets (in TB) as a function of the number of accesses in the last 26 weeks. Top: physical volume, bottom: logical volume 25

Figure 5-4: Volume of accessed datasets (in) TB as a function of the number of accesses in the last 52 weeks. Top: physical volume, bottom: logical volume 26

List of Tables

Table 1-1: LHCb estimated resource needs for 2023 (as of February 2022)....................1

Table 1-2: Site 2023 pledges for LHCb. 1

Table 3-1: Average CPU power provided to LHCb during Jan-Dec 2023 (Tier0 + Tier1s). 6

Table 3-2: Average CPU power provided to LHCb from Jan 1st until October 31st, 2023, in WLCG sites other than the Tier0 and Tier1s. 7

Table 3-3: number of simulated events, stored events, their ratio, total CPU work, CPU work per event and logical file size (LFS), in recent calendar years. Only simulations corresponding to the Run1+Run2 data taking years are considered. The CPU work is taken from the DIRAC accounting. 15

Table 4-1: Disk Storage resource usage as of December 31st, 2023, for the Tier0 and Tier1 centers. The top row is taken from the LHCb accounting, the other ones (used, available and installed capacity) are taken from the WLCG Storage Space Accounting (WSSA) tool. The 2023 pledges are shown in the last row. (*) CERN and RAL do not publish the tape buffer information in WSSA. The additional Tier0 buffer mentioned in the text is not included in the table..17

Table 4-2: Available and used disk resources at Tier-2D sites on December 31st, 2023, taken from WSSA. Sites marked with § do not publish to WSSA; the information from DIRAC is reported instead. 21

Table A-0-1: Summary of the scrutinized, pledged, deployed resources in the 2023 WLCG year. The CPU resources used in the 2023 calendar year, and the occupancy of storage resources on 31/12/2023, are also reported. As CERN and RAL do not publish their tape buffer information, the Tier0 and Tier1 disk occupancies reported in the table below are lower than the real ones. 28

1. Introduction

As part of the WLCG resource review process, experiments are periodically asked to justify the usage of computing resources that have been allocated to them. The requests for the 2023 period were presented in LHCb-PUB-2022-010. Table 1-1 shows the requests to WLCG. For CPU, an additional average power (over the year) of 100kHS06 was deemed to be available from the HLT and other opportunistic resources.

2023 needs	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier0	215	30.3	91
Tier1	705	60.5	157
Tier2	390	11.6	n/a
Total WLCG	**1310**	**102.5**	**248**

Table 1-1: LHCb estimated resource needs for 2023 (as of February 2022).

The last update to the pledge of computing resources for LHC from the different sites can be found in http://wlcg-cric.cern.ch/core/vopledgereq/listcomp/. The LHCb numbers for 2023 are summarized in Table 1-2.

2023 pledges	CPU (kHS06)	Disk (PB)	Tape (PB)
Tier 0	215	30.3	91
Tier 1	598	54.7	134
Tier 2	434	7.9	
Total WLCG	**1247**	**92.9**	**225**
Non WLCG	**50**		

Table 1-2: Site 2023 pledges for LHCb.

The present document covers the resource usage by LHCb in the period January 1st – December 31st, 2023. It is organized as follows: section 2 presents the computing activities; section 3 shows the usage of CPU resources; section 4 reports on the usage of storage resources. Section 5 shows the results of a study on data popularity. A summary is given in Section 6. A schematic view of the resource utilization is given in Appendix A, while answers and comments to the C-RSG remarks from their Autumn 2023 report are reported in a separate document [LHCb-PUB-2024-002].

2. Computing activities during 2023

The usage of offline computing resources involved: (a) the production of simulated events; (b) running user jobs; (c) centralized production of ntuples for analysis working groups; (d) centralized processing of the proton and heavy ion collision datasets collected in 2022 and 2023; (e) an incremental stripping cycle of the Run2 proton collision datasets.

The first three workflows run continuously throughout the year. The fourth and fifth were concentrated towards the end of the year, with a tail extending to January 2024.

As in previous years, LHCb continued to make use of opportunistic resources, that are not pledged to WLCG, but that significantly contributed to the overall usage.

When preparing the 2023 computing requests, it had been assumed that 2023 would have been a full data-taking year for LHCb. However, an incident in the LHC vacuum system close to the LHCb VELO detector implied that the VELO had to be kept open throughout 2023, with impact on the LHCb detector acceptance and somewhat degraded momentum and invariant mass resolutions. For the same reason, the instantaneous luminosity was significantly lower than the nominal value. The initial data taking period in spring 2023 was mainly devoted to commission the various LHCb subdetectors, resulting in a very small fraction of proton collision data taken under stable conditions. An unforeseen LHC incident in mid-July 2023 caused the anticipated end of the 2023 proton collision run. The corresponding data volumes were therefore much smaller than anticipated, enough to be kept on the online disk buffers; this gave flexibility for improving the full-software trigger dedicated to this dataset.

After the HLT2 processing at the online farm, the 2023 proton collision dataset was then transferred to the offline system in Autumn 2023 for further processing ("sprucing") and finally distributed to the grid.

LHC resumed operations in September 2023, setting up and providing heavy ion collisions. LHCb took part in this activity, collecting data that were subsequently triggered upon, spruced, and distributed offline to the grid.

An offline HLT2 reprocessing of the 2022 proton collision dataset was performed as well, with subsequent sprucing and distribution of the output to the grid.

Finally, an incremental stripping of the Run2 proton collision dataset was performed, to further exploit the physics potential of those data.

In summary, many activities related to processing of Run2 and Run3 data were performed over a short period of time. However, the Run3 datasets were small, and the large increment in storage that was foreseen for 2023 did not materialize; the storage occupancy was therefore below the expected level, with the notable exception of the Tier0.

Disk storage at the Tier0 was under stress and saturated for two reasons:

1. The accumulation of special datasets that were foreseen to be used temporarily but that, following the various incidents and special set-ups that were implemented for the 2022 and 2023 data taking, became precious for a detailed understanding of the performance of the LHCb detector, its calibration, and its alignment.

2. An LHCb oversight of the operations of the Tier0 tape system CTA, which demands a disk buffer, where data are stored before being moved to tape, and which was not explicitly considered in the storage requirements.

Due to these two issues, the pledged disk space at the Tier0 of 30.3 PB was completely full in summer and autumn 2023. We acknowledge the help of CERN/IT in alleviating the peak demand by lending LHCb additional disk space.

Nevertheless, this additional disk space remains necessary to accommodate the special datasets until the end of Run3, and to guarantee a proper disk buffer in front of CTA. A corresponding provision at the Tier0 has therefore been included in the LHCb computing model and associated needs for 2024 and 2025.

The CPU work estimated for sprucing the 2022 and 2023 data, and the production of simulated events corresponding to Run3 data taking, were also much lower than initially foreseen. The bulk of simulation production of Run1+Run2 conditions is ended, with a long tail that still used the most significant fraction of computing resources, for a total usage of CPU well above the pledges.

3. Usage of CPU Resources

Computing activities (see Figure 3-1) were dominated by Monte Carlo production. Figure 3-1 also shows a continuous contribution due to user jobs and working group productions. The average CPU power (2.17MHepSpec23, as measured by DIRAC) is 6% lower than that measured in 2022 (2.31MHepSpec23). This is consistent with the decreasing pressure to generate Run1+Run2 simulated datasets, and the decreased needs for processing and simulating Run3 datasets.

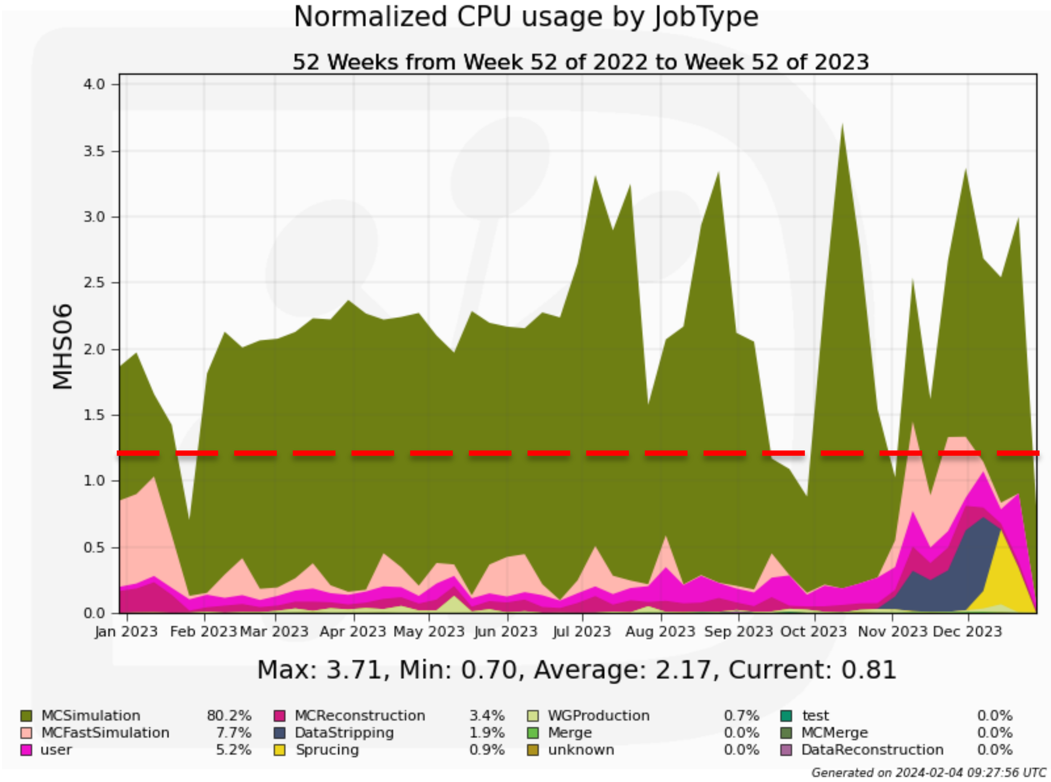


Figure 3-1: Summary of LHCb computing activities (used CPU power) from Jan 1st to December 31st, 2023. The line represents a weighted average of the 2022 and 2023 pledges.

The jobs were executed in the various centers, according to the computing resources available to LHCb (Figure 3-2). The Tier0 and seven Tier1s contribute about 63%, the rest is due to Tier2s, with significant amounts from Tier-2D with disk, and unpledged resources.

The LHCb Online HLT farm was exclusively used by the software trigger system to process the datasets collected by the LHCb detector.

The CPU usage was significantly higher than the requests and the pledges, due to the requests for simulated samples from the physics working groups. A significant fraction of these requests was produced with fast simulation techniques.

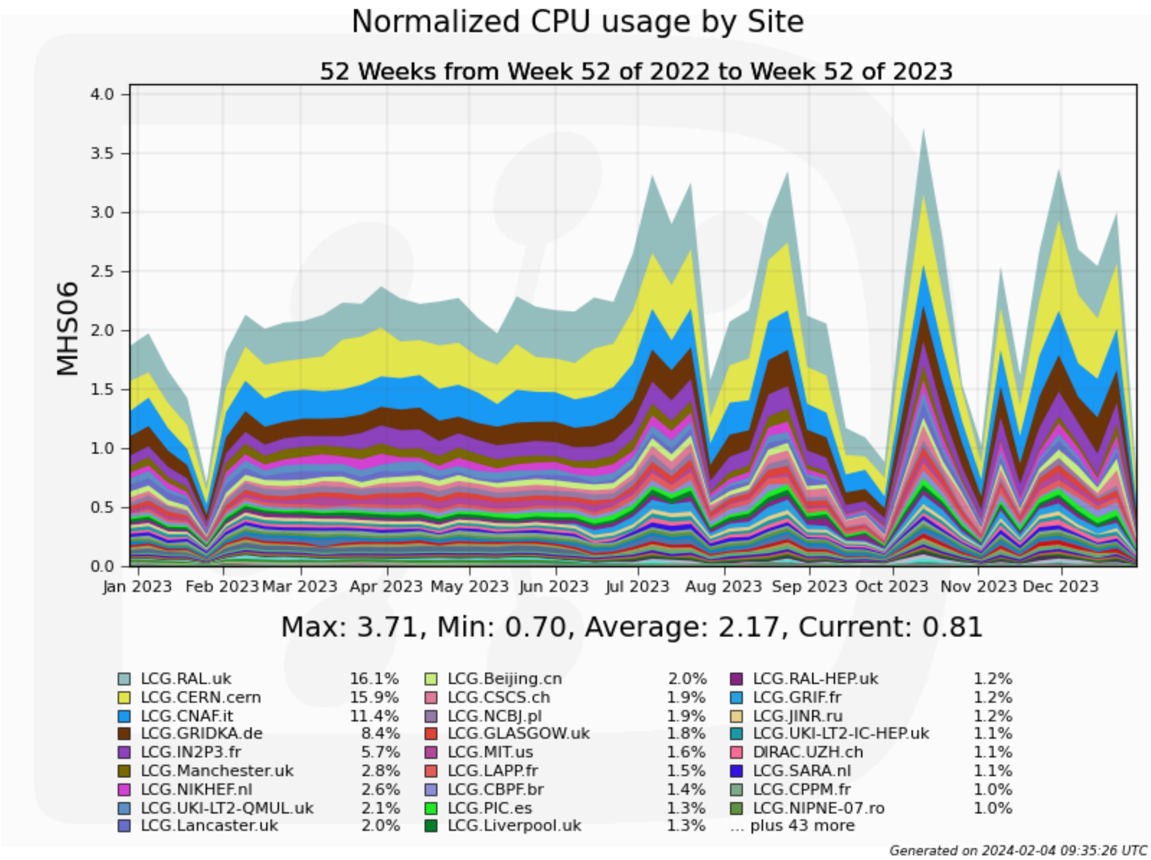


Figure 3-2: Summary of LHCb site usage from Jan 1st until December 31st, 2023.

3.1. WLCG Accounting

The usage of WLCG CPU resources by LHCb is obtained from the different views provided by the WLCG Accounting Utility (WAU) portal. The CPU usage is presented in Figure 3-3 for the Tier1s and in Figure 3-4 for Tier2s[1]. The same data is presented in tabular form in Table 3-1 and Table 3-2.

The average power used at Tier0+Tier1 sites is about 16% higher than the pledges. The average power used at Tier2s is 20% higher than the pledges.

The average CPU power accounted for by WLCG (including Tier0, Tier1 and Tier2 sites) amounts to 1402 kHS06, to be compared to 1310 kHS06 estimated needs quoted in Table 1-1. Additional computing power was used at non-WLCG sites, for an estimated contribution of about 26 kHS06 on average. This number is taken from the DIRAC

[1] Including WLCG Tier2 sites that are not pledging resources to LHCb but are accepting LHCb jobs.

accounting and rescaled to account for known differences between the EGI and DIRAC systems, see Section 3.4.

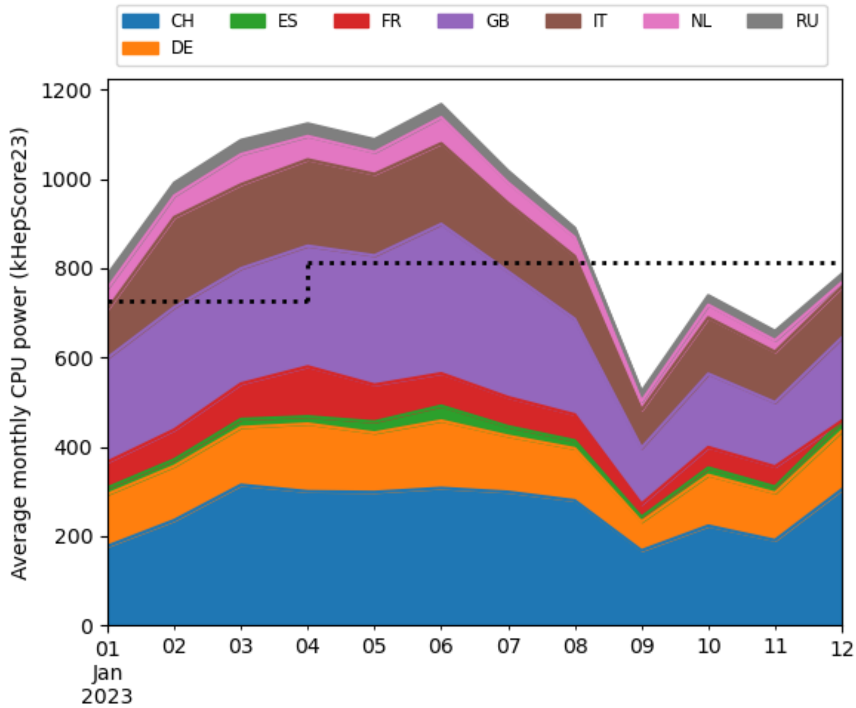


Figure 3-3: Average monthly CPU power provided by the Tier1s (and Tier0) to LHCb from Jan 1st until December 31st, 2023.

<Power>	Used (kHS06)	Pledge (kHS06)
CH-CERN	258.1	208.5
DE-KIT	121.7	114.5
ES-PIC	18.7	24.5
FR-CCIN2P3	64.9	82.3
IT-INFN-CNAF	148.8	110.9
NL-T1	41.2	56.8
RRC-KI-T1	25.6	16.4
UK-T1-RAL	231.5	172.1
Total	**910.5**	**786.0**

Table 3-1: Average CPU power provided to LHCb during Jan-Dec 2023 (Tier0 + Tier1s).

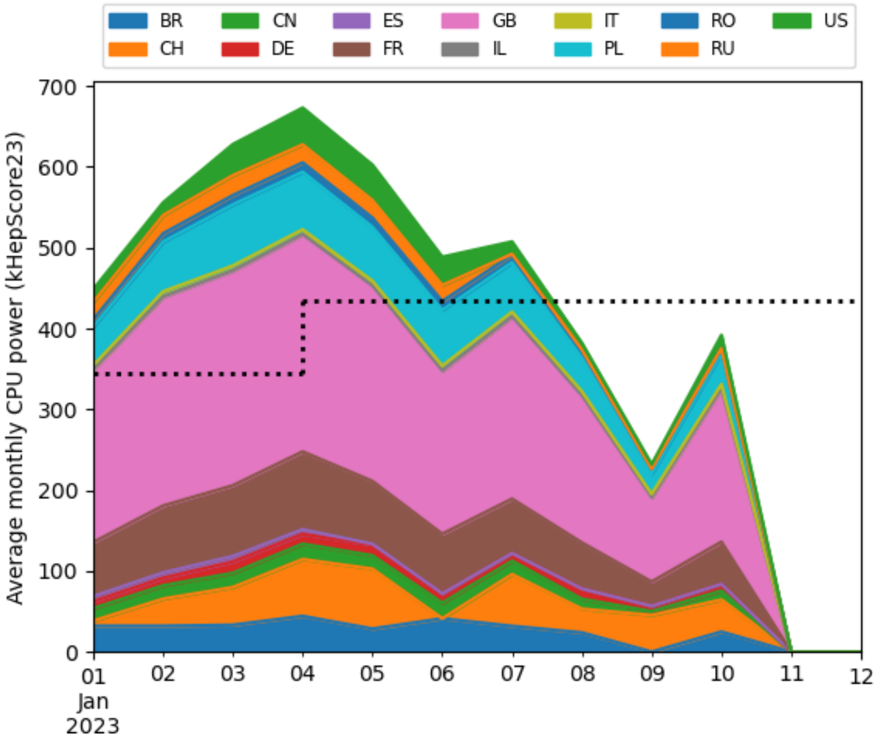


Figure 3-4: Average monthly CPU power provided by the Tier2s to LHCb from Jan 1st until October 31st, 2023.

<Power>	Used (kHS06)	Pledge (kHS06)
China	14.9	28.5
Brazil	29.3	1.5
France	69.1	45.1
Germany	10.6	19.8
Israel	1.0	0.0
Italy	5.3	61.2
Poland	55.7	49.4
Romania	9.5	14.9
Russia	13.6	5.4
Spain	3.4	10.8
Switzerland	40.9	66.5
UK	214.5	68.1
US	24.0	37.5
Total	**491.7**	**408.7**

Table 3-2: Average CPU power provided to LHCb from Jan 1st until October 31st, 2023, in WLCG sites other than the Tier0 and Tier1s.

3.2. LHCb DIRAC CPU Accounting

The sharing of CPU time (in days) is shown in Figure 3-5. The top chart reports the CPU time per country provided by WLCG sites, including the ones not providing a pledge for LHCb. The bottom chart shows the CPU time per site provided by non-WLCG sites. There are sizeable contributions from a few farms (Monash, Sibir, Birmingham and Zurich).

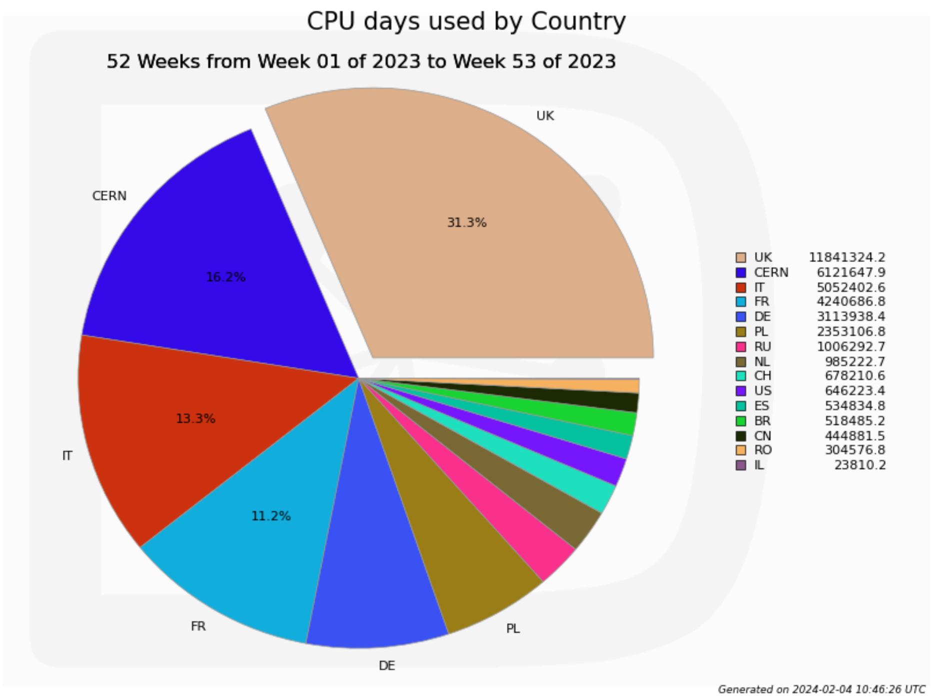
The CPU power used at Tier0/1s is detailed in Figure 3-6. As seen in the top figure, simulation is dominant (96% of the total), with user jobs and working group productions giving the remaining contributions.

The contributions from the Tier2s and other WLCG sites (including sites not pledging any resources to LHCb but excluding the HLT farm) are shown in Figure 3-7. Simulation (97%) dominates the CPU usage. A small fraction of user jobs is also visible in Figure 3-7.

Figure 3-8 shows the number of user and analysis production jobs per site at Tier0 and Tier1s (top left) and outside (top right). The contributions due to each workflow are shown in the bottom plots. Analysis productions are slowly ramping up; it is expected that they will become the dominant method to perform analysis on Run3 data.

Figure 3-9 shows pie charts of the CPU days used at all sites as a function of the final status of the job (top), and as a function of the error (middle) and the activity (bottom) when jobs are not successful. About 98% of all jobs complete successfully. Job failures are generally due to application misconfiguration or excess of CPU or memory usage; a small fraction is due to failures in uploading outputs. By looking at the bottom plot of Figure 3-9, one can see that failed jobs are due to simulation jobs in about half of the cases, the remainder being due to user jobs and analysis productions.

Most stalled jobs were temporary bursts, due to misconfigured Monte-Carlo productions. The remaining stalled jobs were typically due to failures in getting the proper size of a batch queue, or in evaluating the CPU-work needed for an event (and thus over evaluating the number of events the jobs can run).



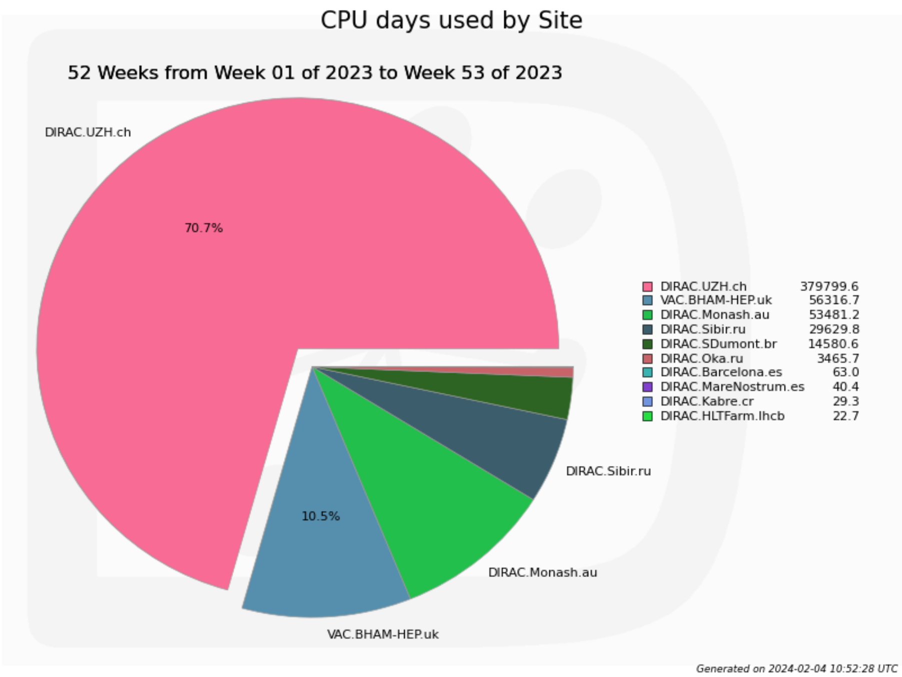


Figure 3-5: CPU time consumed by LHCb from Jan 1st until December 31st, 2023 (in days), for WLCG sites, excluding the HLT farm (top, per country), and non-WLCG sites, including the HLT farm (bottom, per site).

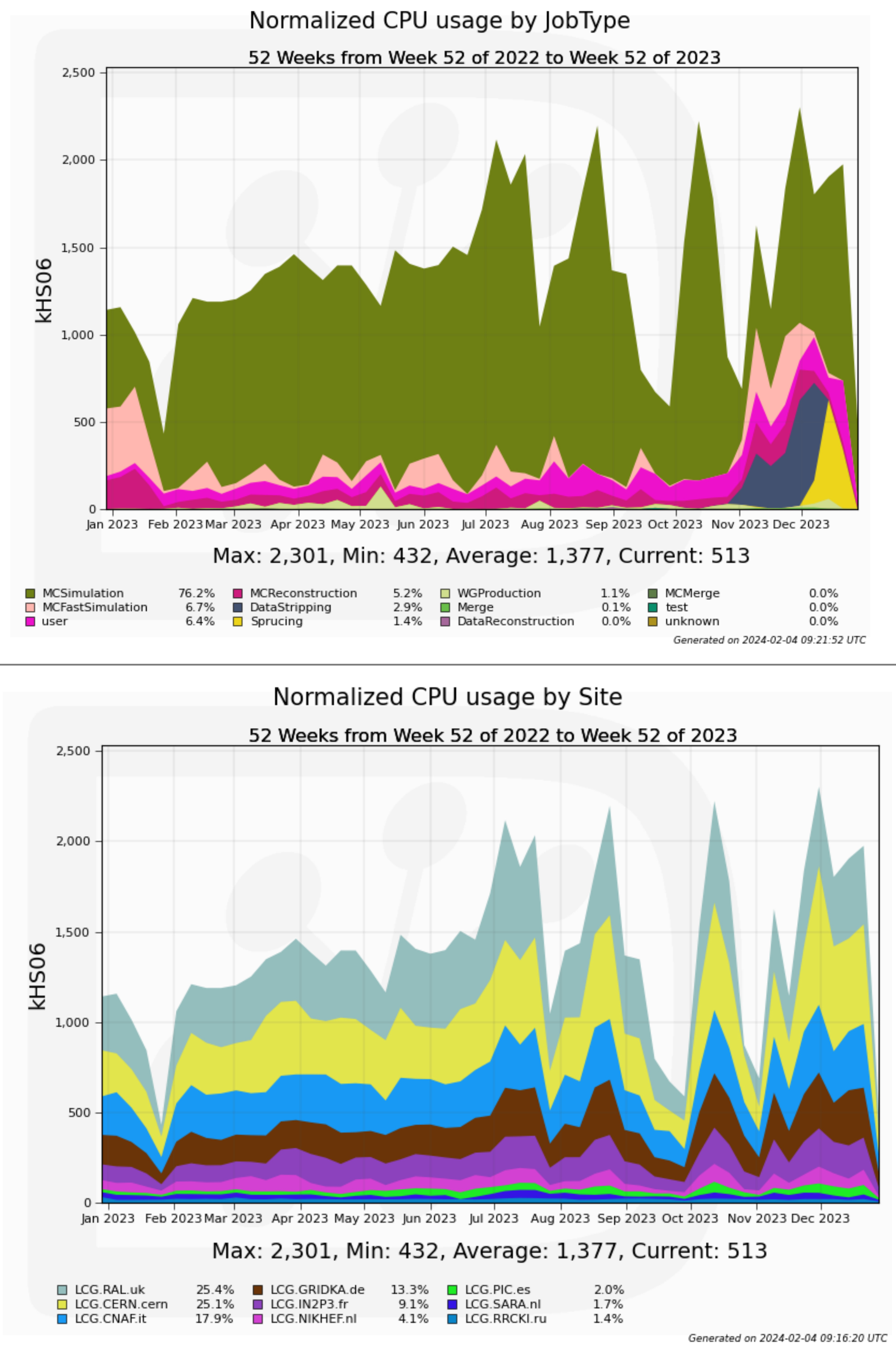
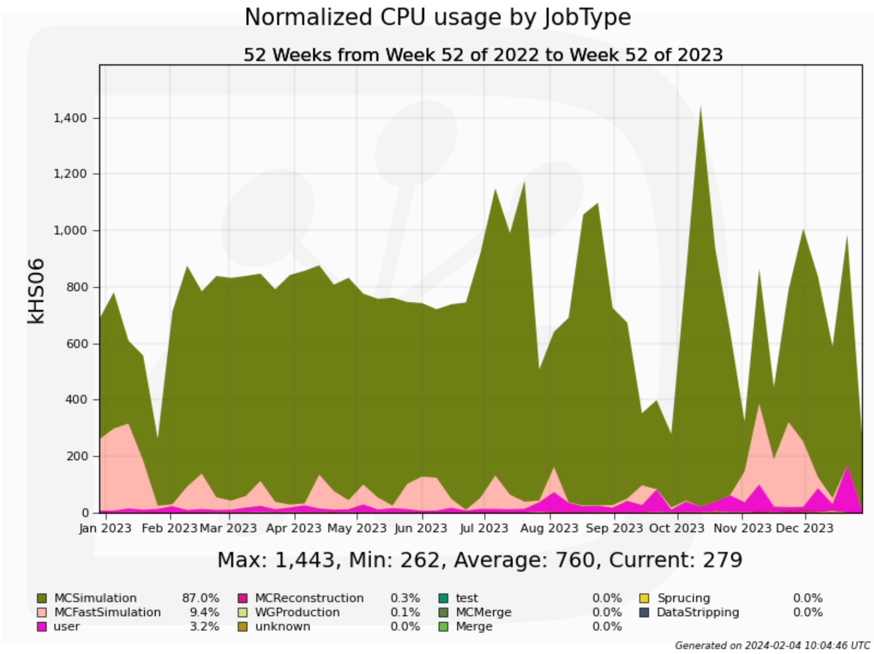


Figure 3-6: Usage of LHCb Tier0/1s during Jan 1st – Dec 31st, 2023. The top plot shows the used CPU power as a function of the different activities, while the bottom plot shows the contributions from the different sites. User jobs and analysis productions (shown in magenta and light green in the top plot) are further detailed in Figure 3-8.



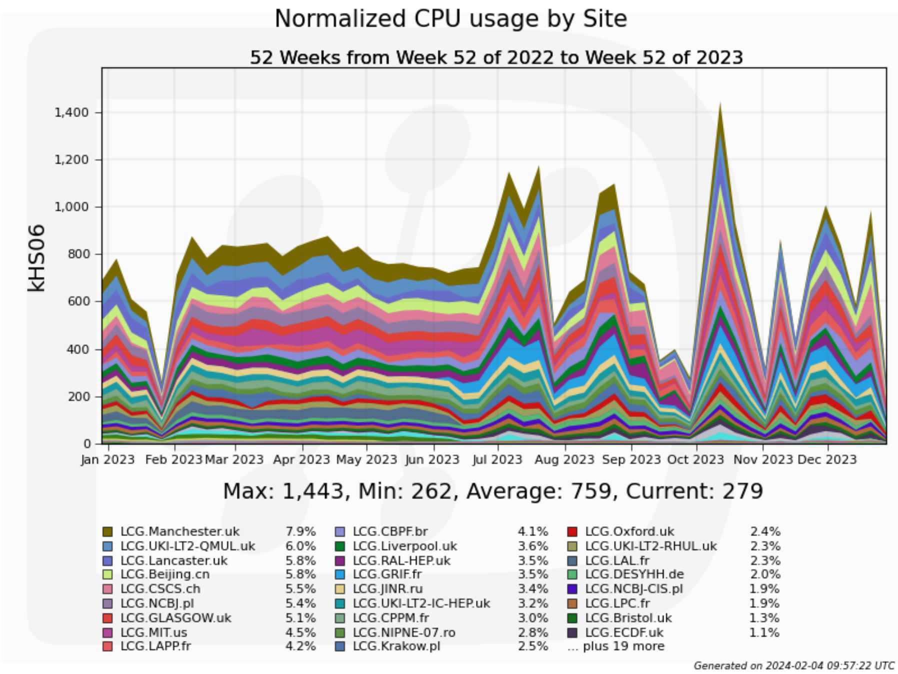


Figure 3-7: Usage of resources in WLCG sites other than Tier0 and Tier1s, during Jan 1st—Dec 31st, 2023. The top plot shows the used CPU power as a function of the different activities, while the bottom plot shows the contributions from the different sites. User jobs (shown in magenta in the top plot) are further detailed in Figure 3-8.

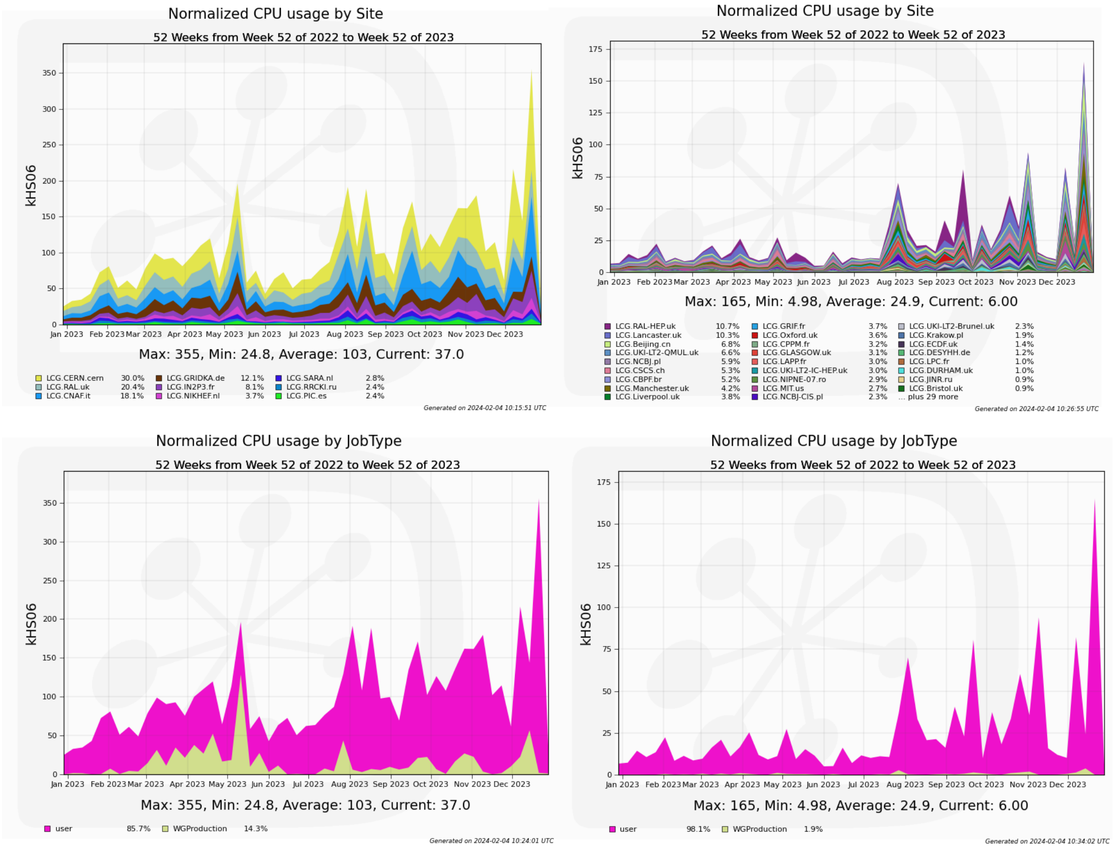


Figure 3-8: Running user and analysis production jobs on Tier0 and Tier1 sites (top left), and outside (top right) during Jan 1st – Dec 31st, 2023. The two lower plot show the contribution of the two workflows (user jobs in magenta, analysis productions in light green)

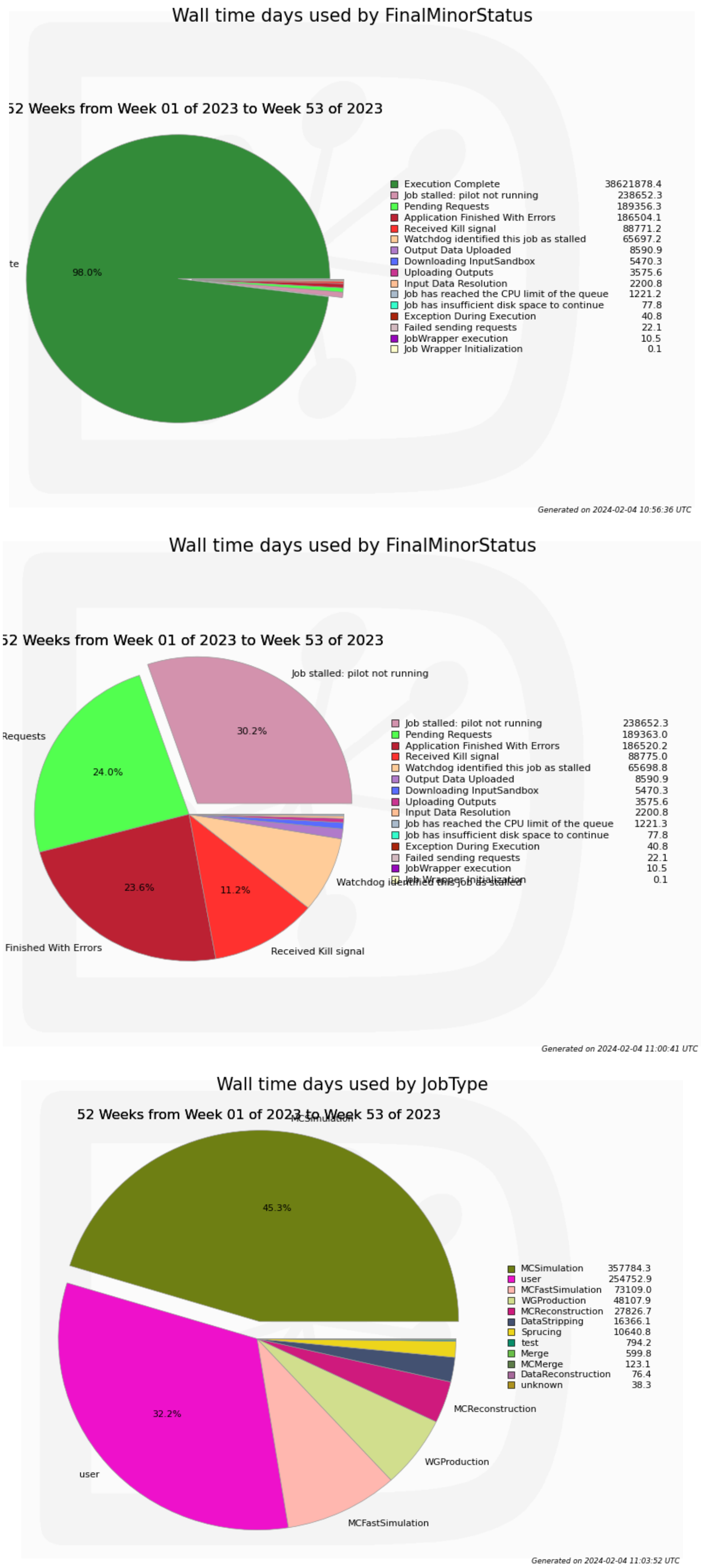


Figure 3-9: CPU time as a function of the job final status for all jobs (top); for failed jobs, the middle and bottom plots show the error status and activity, respectively.

3.3. Usage of Fast Simulations and filtered productions

Fast and detailed simulations accounted in almost equal proportions in the number of events produced in 2023 (Figure 3-10, left). The number of produced events per day as a function of time is shown in (Figure 3-10, right). The number of simulated events and total CPU work in 2023 is lower than previous years (Figure 3-11, Table 3-3). This is due to the gradual completion of Run1+Run2 simulation productions, matching the assumptions in the computing model, and to the absence of significant Run3 simulation production.

To save storage resources, only events passing analysis-dependent selection criteria are persisted on storage. The fraction of persisted events in 2023 is 4.1% (Table 3-3).

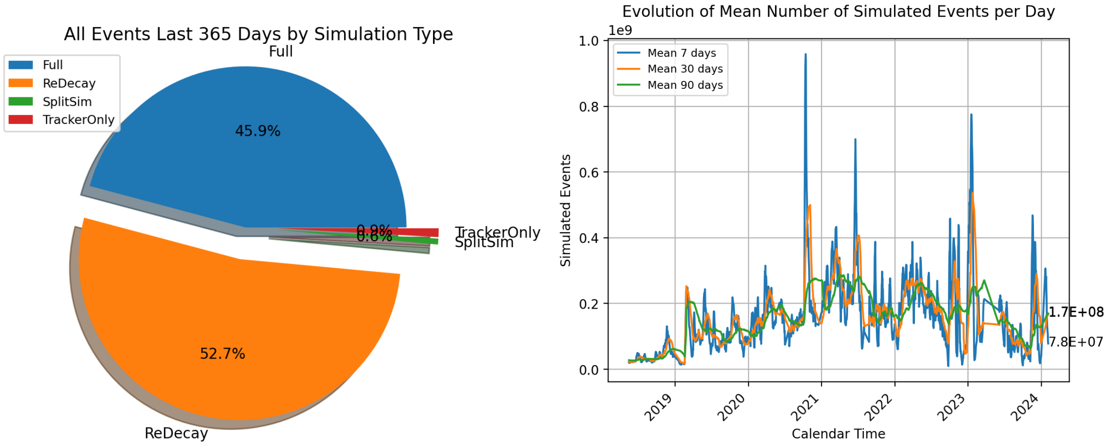


Figure 3-10: Left: number of events produced in 2023 by simulation type; right: number of events produced per day, with rolling averages over 7, 30, 90 days.

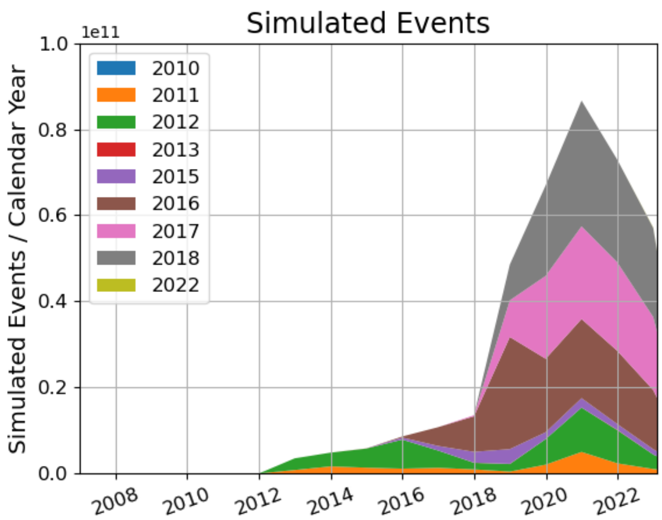


Figure 3-11: Number of events simulated as a function of calendar year. Each color corresponds to the simulation of a given data-taking year.

Year	Simulated events (10^9)	Stored events (10^9)	Ratio	CPU work kHS06.y	CPU per event kHS06.s	LFS TB
2017	10.3	4.2	40.3%	817	2.50	640
2018	12.0	3.0	25.3%	1009	2.65	550
2019	45.0	6.9	15.2%	1290	0.90	1110
2020	67.0	16.8	31.7%	1357	0.81	2010
2021	80.0	11.1	13.9%	1815	0.72	2030
2022	78.4	3.2	4.1%	2243	0.98	2490
2023	55.4	2.2	4.1%	1981	1.13	1450

Table 3-3: number of simulated events, stored events, their ratio, total CPU work, CPU work per event and logical file size (LFS), in recent calendar years. Only simulations corresponding to the Run1+Run2 data taking years are considered. The CPU work is taken from the DIRAC accounting.

3.4. Differences between WLCG and DIRAC accounting

Due to the different approaches used to convert the used CPU seconds in CPU work, the WLCG and DIRAC accounting systems give slightly different measurements. The former uses a factor that is given on a site-by-site basis and that generally corresponds to the average power (in HepSpec23) of a worker node. The latter converts the used CPU seconds in CPU work by running in every single job a reference workload to estimate the power of the node where the job is executed.

For consistency reasons, the results of the WLCG accounting are taken as reference for the Tier0, Tier1 and Tier2. The used CPU from sites not monitored in WLCG is determined by taking the DIRAC measurements and rescaling them to account for the mismatch between the two systems.

4. Usage of Storage resources

4.1. Tape storage

Due to the Run3 datasets, tape storage grew by about 17 PB. The total tape occupancy as of December 31st, 2022, is 93.6 PB, 49.9 PB of which are used for RAW data, 14.6 PB for RDST, 29.1 PB for archived data. This is 65% lower than the original request of 248PB, which was made under very different assumptions for data taking in 2022 and 2023.

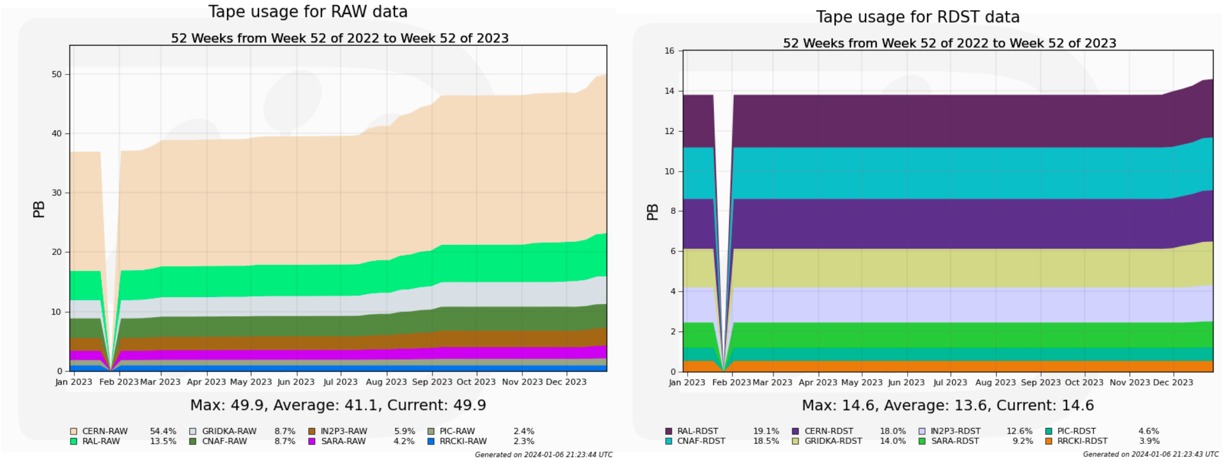


Figure 4-1: Tape space occupancy for RAW (left) and RDST (right) data, January 1st – December 31st, 2023. The glitch at the end of January in this and in the following plots corresponds to an outage of the LHCb accounting system and it is not significant.

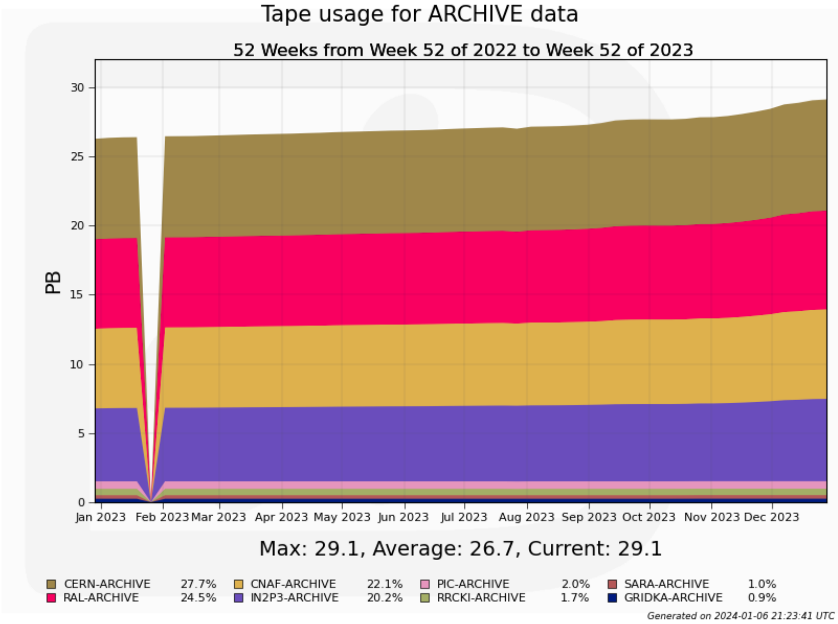


Figure 4-2: Tape occupancy for derived data, January 1st – December 31st, 2023. The glitch at the end of January corresponds to an outage of the LHCb accounting system and it is not significant.

4.2. Disk at Tier0, Tier1s and Tier2s

The evolution of Disk usage during 2023 at the Tier0 and Tier1s is presented in Figure 4-3, separately for real data DST and MC DST, Figure 4-4 for disk buffer space, and Figure 4-5 for user disk.

There was a clean-up of both real data and MC DST disk space in summer 2023. For real data, 0.5PB were recovered by deleting a previous processing pass of Run2 data. For simulation, datasets that have not been accessed in the previous two years were deleted, thus recovering 2.5 PB. The deleted datasets are archived on tape, and they can be staged back on disk if users need to process them; this occurred only in a handful of cases, with minimal impact on data management and operations. The increase of data DST disk space towards the end of the year corresponds largely to the incremental stripping of Run2 data and, to a lesser extent, to the sprucing of Run3 proton and heavy ion collisions datasets.

The disk buffer space is normally used for storing temporary files such as RAW and RDST until they are stripped or DST files before they are merged into large files. There are several increasing steps in the usage of buffer space, corresponding to the distribution of 2022 and 2023 proton and heavy ion collision datasets; in the last three months of 2023, Run2 proton collision datasets were staged from tape in preparation of their incremental stripping. A maximum of 12.4 PB was achieved at the beginning of December, with a decrease afterwards, corresponding to the deletion of input data after they have been processed.

The space used at the Tier0 and Tier1 sites increased by 11.9PB, almost entirely due to the Tier0 buffer area mentioned in Section 2, and not shown in the plots.

Table 4-1 shows the situation of disk storage resources at CERN and Tier1s, as well as at each Tier1 site, as of December 31st, 2023. The used space includes derived data, i.e. DST and micro-DST of both real and simulated data, and space reserved for users. The latter consists of two contributions: a distributed one at the Tier0 and Tier1 sites (1.8PB in total, 1.4 of which are used, see Figure 4-5), the EOS space reserved for physics analysis working groups at CERN (1.7PB in total, 1.34PB of which are used). The additional EOS buffer space mentioned in Section 2 adds another 11.9 PB in total at the Tier0, 6.7 PB of which are used.

Disk (PB)	CERN	Tier1s	CNAF	GRIDKA	IN2P3	PIC	RAL	RRCKI	SARA
LHCb accounting	**11.0**	**31.8**	**6.3**	**5.8**	**4.9**	**1.6**	**8.4**	**2.0**	**3.1**
WSSA disk used	11.5	32.8	5.8	5.9	5.0	1.7	9.7	1.7	3.1
WSSA disk free	4.1	19.8	3.8	3.6	3.1	0.6	6.0	0.8	1.9
WSSA tape buffer (used+free)[3]	(*)	2.4	1.8	0.3	0.1	0.05	(*)	0.2	0.05
WSSA total[2]	**15.5**	**55.0**	**11.4**	**9.8**	**8.1**	**2.3**	**15.7**	**2.7**	**5.0**
Pledge 2022	***30.3***	***54.7***	***11.6***	***10.1***	***7.7***	***2.4***	***15.7***	***2.3***	***4.9***

Table 4-1: Disk Storage resource usage as of December 31st, 2023, for the Tier0 and Tier1 centers. The top row is taken from the LHCb accounting, the other ones (used, available and installed capacity) are taken from the WLCG Storage Space Accounting (WSSA) tool. The 2023 pledges are shown in the last row. (*) CERN and RAL do not publish the tape buffer information in WSSA. The additional Tier0 buffer mentioned in the text is not included in the table.

[2] This total includes disk visible in WSSA for tape cache. It does not include invisible disk pools for dCache (stage and read pools).

The *WSSA disk used* and *WSSA disk free* information concerns only permanent disk storage. The first two lines show a good agreement between what the site reports and what the LHCb accounting (first line) reports, except for RAL[3]. The sum of the Tier0 and Tier1s 2021 pledges amount to 85 PB. The available disk space is 70.5 PB in total, 44.3 PB of which are used to store real and simulated datasets, and user data. The remainder is allocated for tape buffer, 2.4PB of which are accounted in WSSA. The Tier0 temporary buffer of 11.9 PB previously mentioned is not accounted for in WSSA. Therefore, the total space available to LHCb at the end of 2023 at the CERN Tier0 is 15.5+11.9=27.4 PB.

The available disk at Tier-2D is now 7.3 PB, 3.6 PB of which is used. The disk pledge at Tier-2D sites is 7.9 PB. Table 4-2 shows the situation of the disk space at the Tier-2D sites. Figure 4-6 shows the disk usage at Tier-2D sites. A similar pattern to that observed for the Tier0 and Tier1 sites is observed: clean-up of real and simulated datasets, and increases due to freshly produced data.

In conclusion, at the end of the 2023 calendar year, LHCb was using 62.2 PB of disk space, 20% lower than the total available disk space of 78.2 PB and 33% lower than the pledge of 92.9PB. In the first quarter of 2024, the disk space occupancy will increase, due to the continuous production of simulated events and to the tail of the Run2 and Run3 processing activities. LHCb will not saturate the disk space allocated in the 2023 WLCG year.

[3] For CERN, the EOS working group space is included in the WSSA report, but not in the LHCb report. For RAL, the difference is due to dark data i.e., data that are stored and reported in WSSA but that are not in the LHCb bookkeeping catalog and therefore not accounted for in the LHCb accounting.

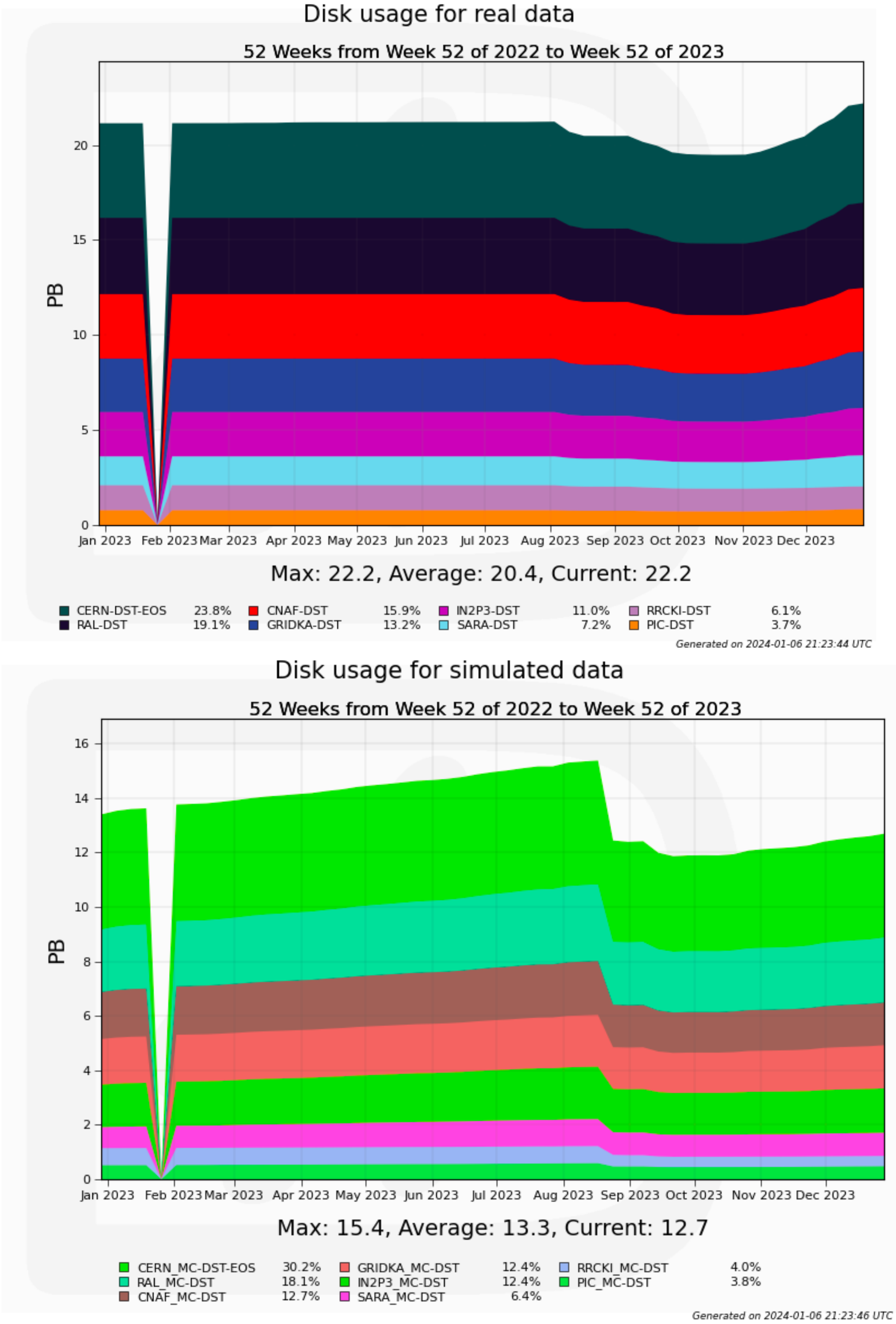


Figure 4-3: Usage of Disk resources at CERN and Tier1s from Jan 1st to December 31st, 2023. Real data and simulated data are shown in the top and bottom figures.

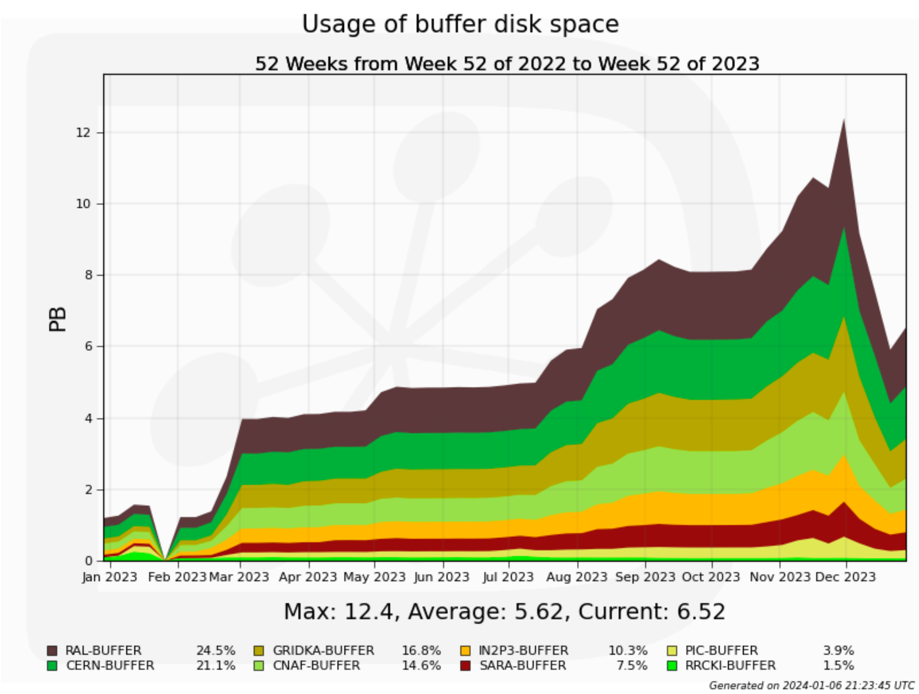


Figure 4-4: Usage of LHCb buffer disk from January 1st to December 31st, 2023.

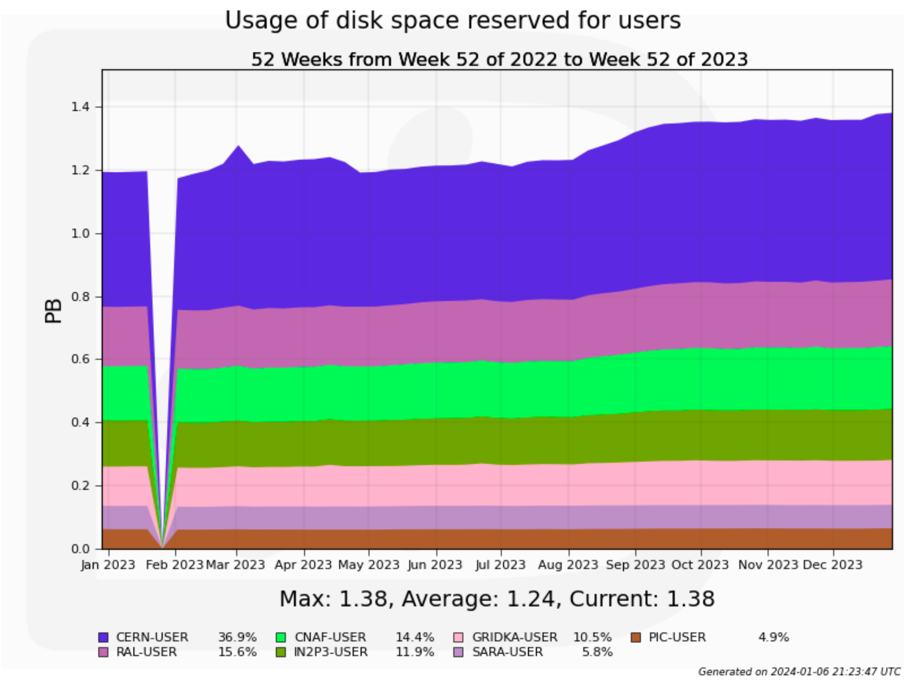


Figure 4-5: Usage of disk resources reserved for Users from January 1st to December 31st, 2023.

Table 4-2: Available and used disk resources at Tier-2D sites on December 31st, 2023, taken from WSSA. Sites marked with § do not publish to WSSA; the information from DIRAC is reported instead.

Site	SRR free disk (TB)	SRR used disk (TB)	SRR total disk (TB)
Beijing (China)	176	199	375
CSCS (Switzerland)	1696	604	2300
GRIF§ (France)	121	271	392
CPPM§ (France)	214	386	600
Glasgow§ (UK)	24	86	110
IHEP Protvino (Russia)	144	144	288
Liverpool§ (UK)	99	213	312
Manchester (UK)	92	218	310
NCBJ§ (Poland)	156	504	660
NIPNE§ (Romania)	320	80	400
RAL-HEP (UK)	194	326	520
UKI-QMUL (UK)	83	217	300
UKI-IC (UK)	358	352	710
Total	**3677**	**3600**	**7277**

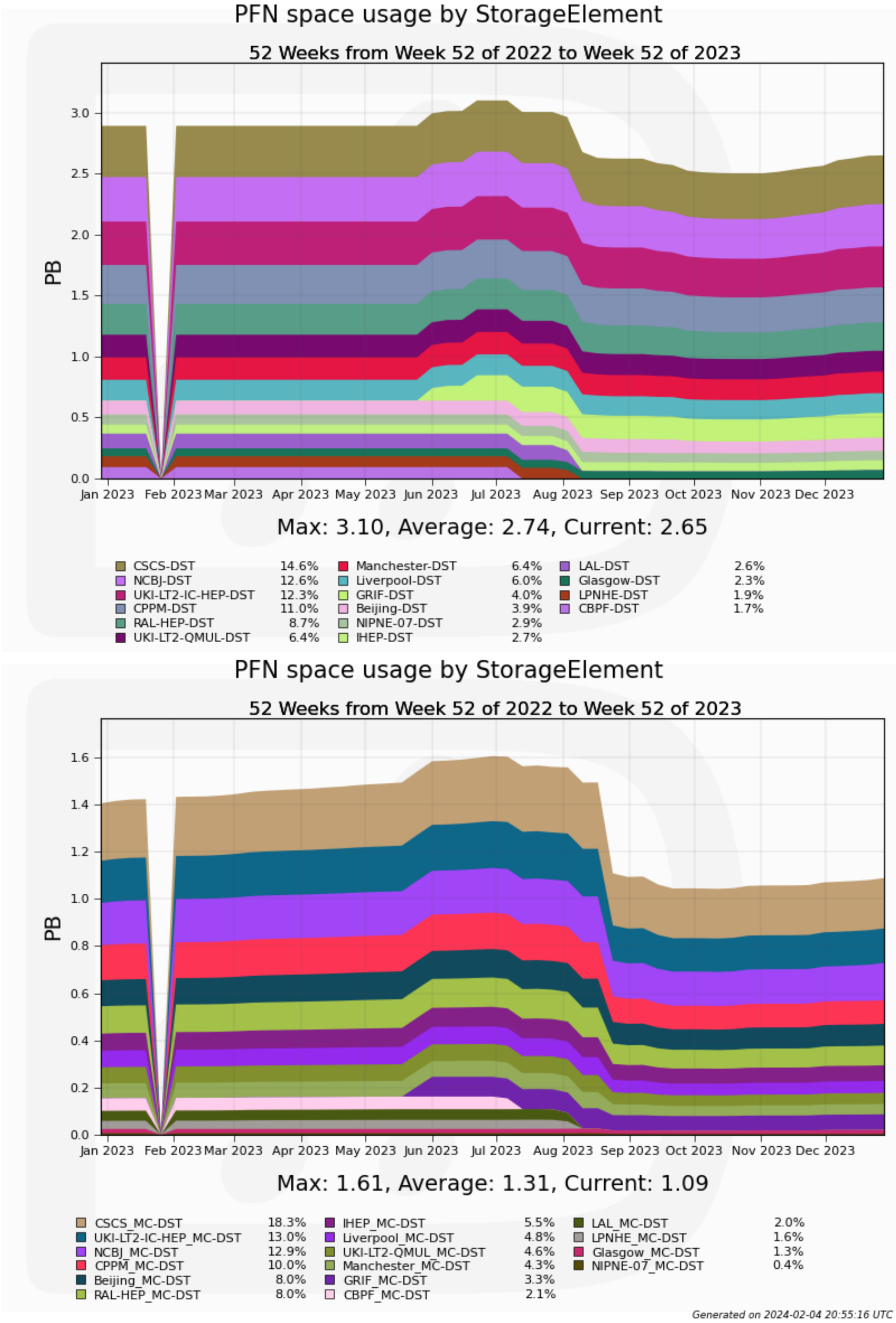


Figure 4-6: Usage of disk space at Tier-2D sites from January 1st to December 31st, 2023, for real data (top) and simulated data (bottom).

5. Data popularity

Dataset usage patterns, routinely monitored in LHCb since some years, are used to recover a significant amount of disk space.

Figure 5-2, Figure 5-3, Figure 5-4 show the physical and logical volume on disk of the accessed datasets as a function of the number of accesses in the last 13, 26, and 52 weeks as measured on January 31st 2024. The number of usages of a dataset over a period is defined as the ratio of the number of files used in that dataset to the total number of files that were not accessed during that period (e.g. 7.1 PB physical volume in the last 52 weeks). The last bin shows the total volume of data that was accessed (e.g., 31.6 PB physical volume in the last 52 weeks)[4]. The total amount of disk space used by the derived LHCb data is 39.5 PB[5]. The datasets accessed at least once in 13, 26 and 52 weeks, occupy respectively 60%, 67% and 80% of the total disk space in use, respectively, somewhat higher than the result of previous popularity studies (Figure 5-1). An analysis in terms of the logical instead of the physical volume of the datasets leads to similar results.

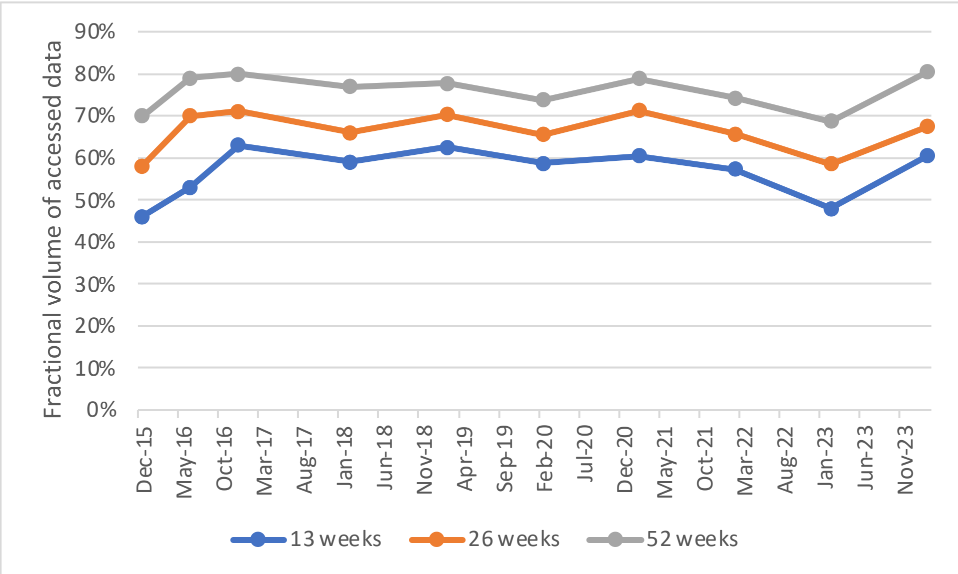


Figure 5-1: Volume (as percentage of total data volume) of datasets accessed at least once in the last 13, 26, 52 weeks, in various periods as reported in this and previous documents.

[4] The volume in each bin is not weighted by the number of accesses.

[5] Since this number is computed by explicitly excluding files that are not usable for physics analysis, e.g. RDST or temporary unmerged files, it is not directly comparable to the used resources mentioned in Section 4.

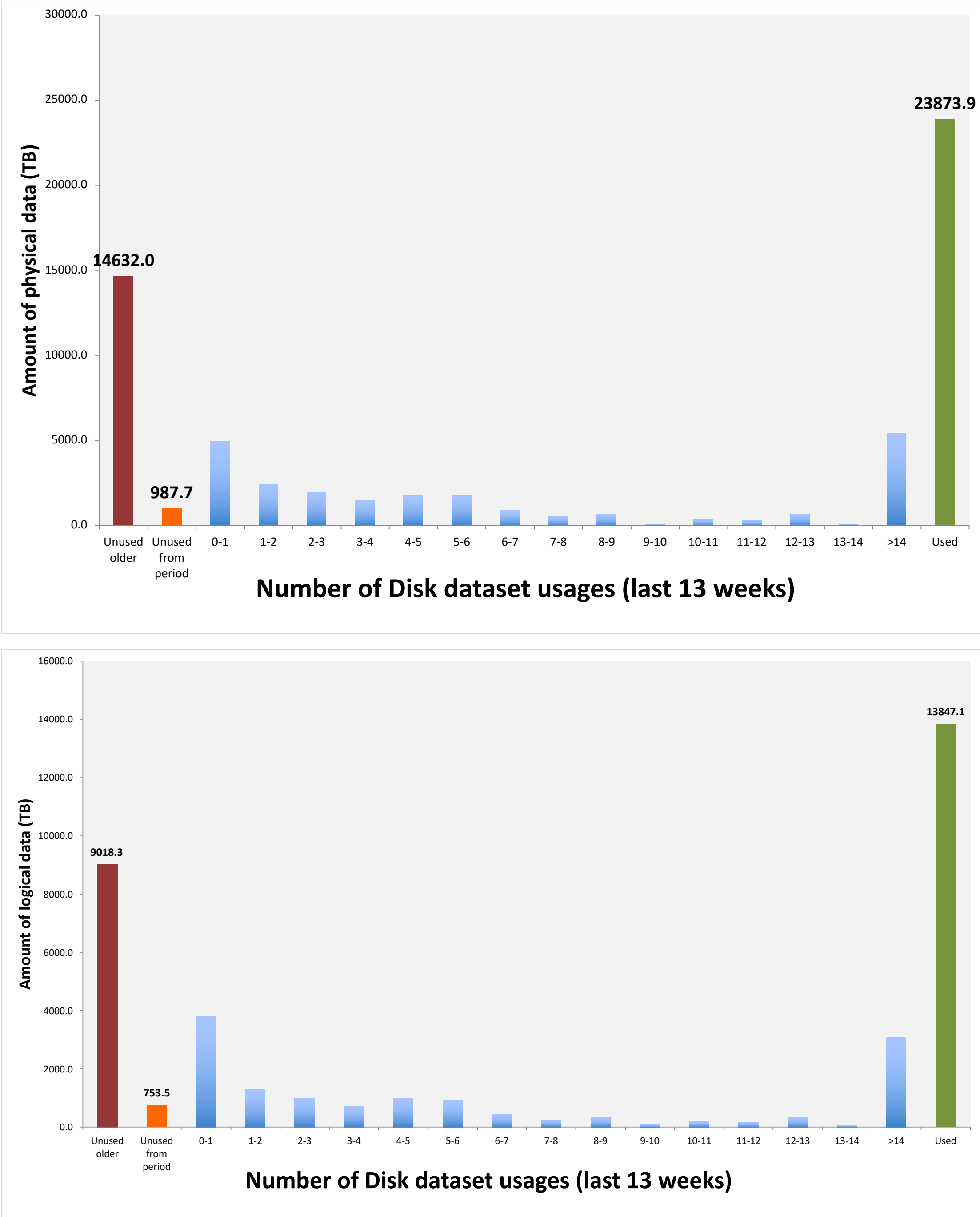


Figure 5-2: Volume of accessed datasets (in TB) as a function of the number of accesses in the last 13 weeks. Top: physical volume, bottom: logical volume

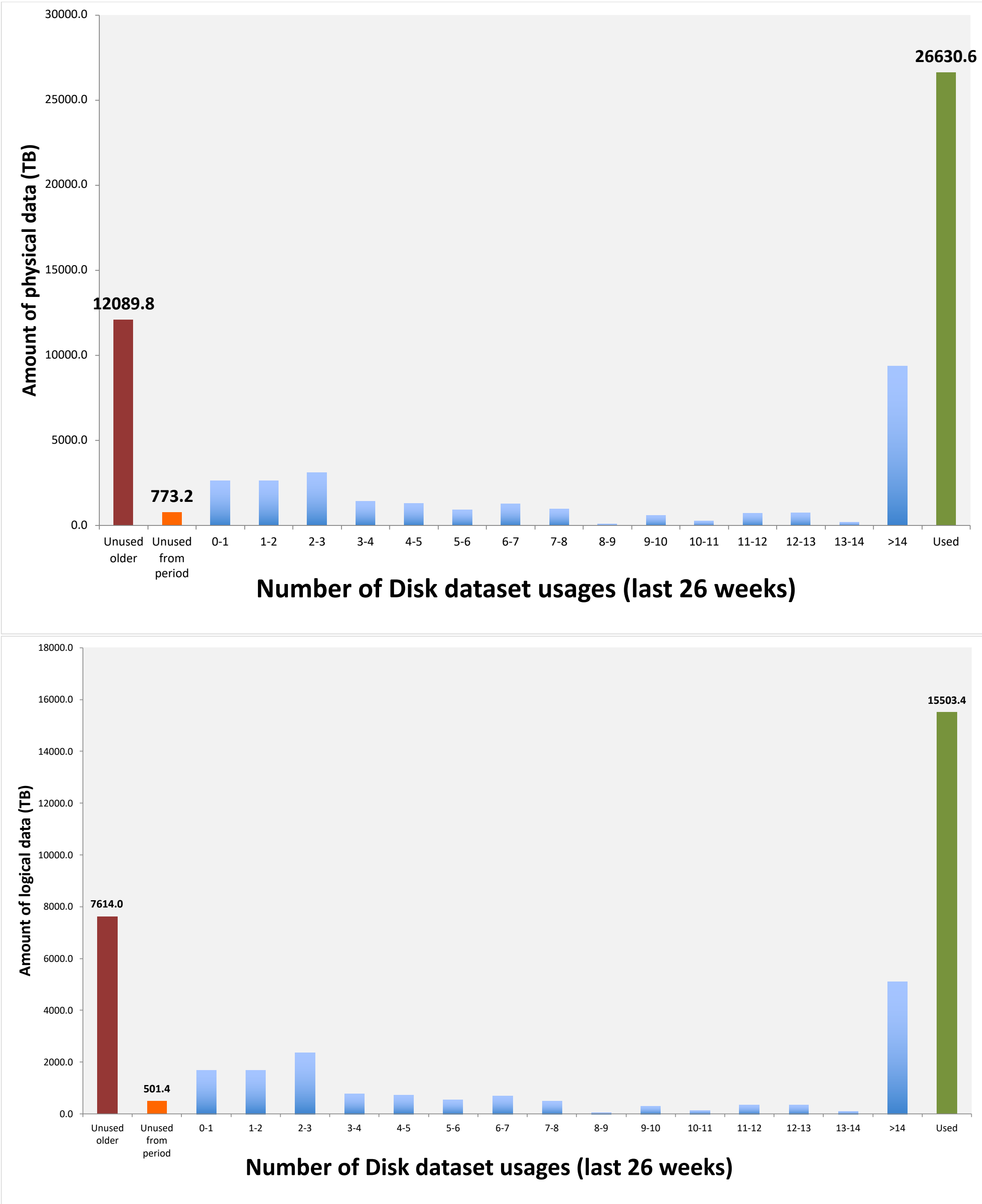


Figure 5-3: Volume of accessed datasets (in TB) as a function of the number of accesses in the last 26 weeks. Top: physical volume, bottom: logical volume

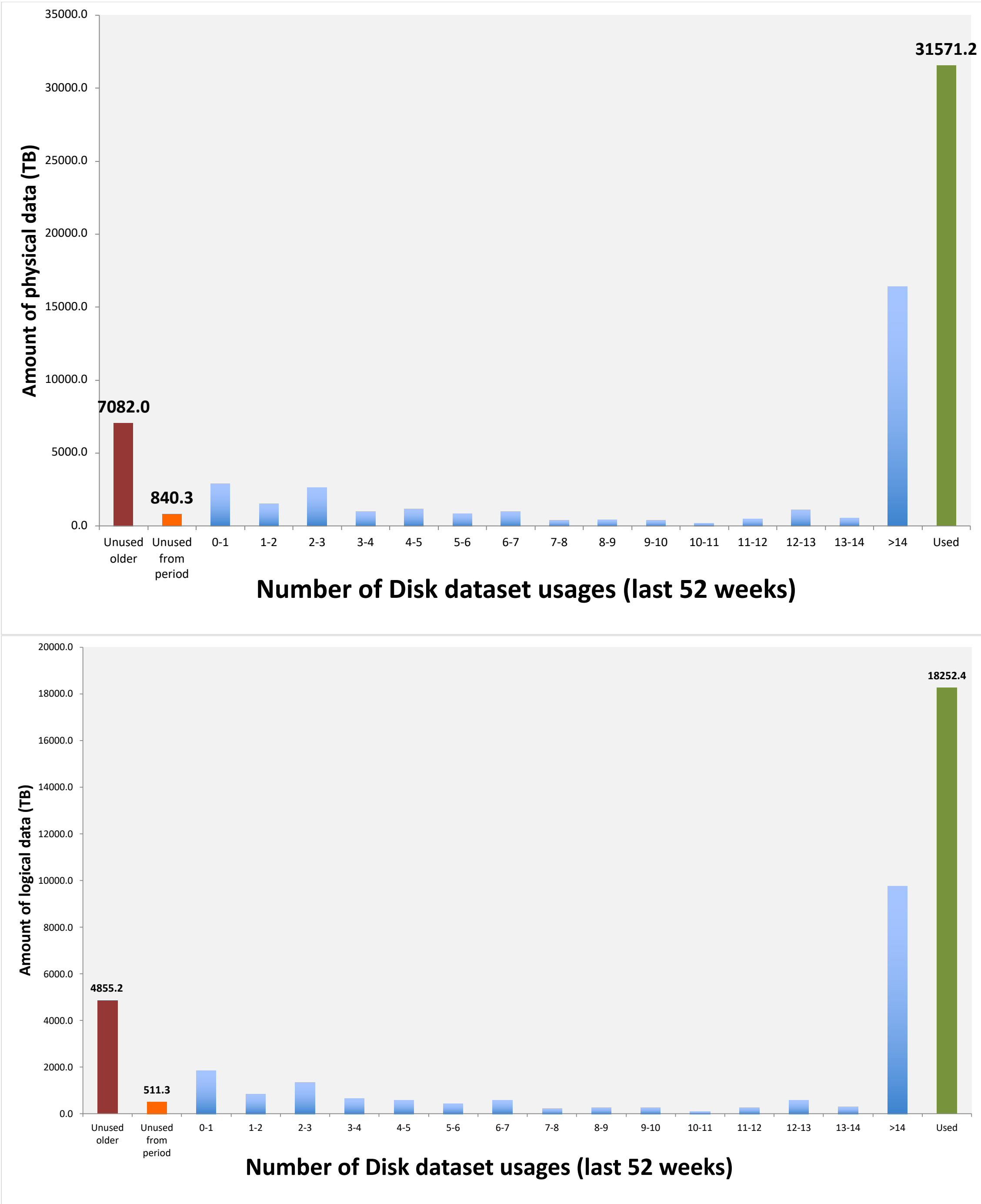


Figure 5-4: Volume of accessed datasets (in) TB as a function of the number of accesses in the last 52 weeks. Top: physical volume, bottom: logical volume

6. Summary

The usage of computing resources in the 2023 calendar year has been in general smooth for LHCb, with the notable exception of disk shortage at the Tier0 during summer, due to unforeseen datasets that had to be kept longer than necessary, and to an LHCb oversight of the disk buffer needed for CTA operations. The operation of the LHCb distributed computing infrastructure continued to be effective. Activities related to the offline processing of Run3 datasets eventually started in Autumn, and an initially unforeseen incremental stripping of Run2 data was performed as well.

The expected increase of storage resources in 2023 did not materialize, due to various issues with the LHC.

Simulation was therefore once more the dominant activity in terms of CPU work. Additional unpledged resources, as well as clouds and on-demand computing resources, were also successfully used. The total number of events produced is lower to respect to last years, as Run1+Run2 simulation productions are phasing out and significant Run3 simulation samples have not been produced yet.

Appendix A

The following table provides a summary of the scrutinized (C-RSG), pledged and deployed resources in the WLCG 2023 year, as well as the used CPU resources in the 2023 calendar year and the storage occupancies on 31/12/2023. Some useful ratios between these quantities are reported as well.

For used CPU, the measurements from the EGI accounting are reported for Tier-0, Tier-1 and Tier-2 sites. For the HLT and other sites that are not accounted in EGI, the DIRAC measurements are taken and rescaled to account for the mismatch between the two systems.

For disk, both used and deployed capacity are taken from WSSA. As CERN and RAL do not publish their tape buffer information, the Tier0 and Tier1 disk occupancies reported in the table below are lower than the real ones.

LHCb		2023						
		Request	Pledge	Pledge/req	Used	Used/CRSG	Deployed capacity	Deployed cap./CRSG
WLCG CPU	Tier-0	215	215	100%	258	120%	215	100%
	Tier-1	705	598	85%	652	92%	598	85%
	Tier-2	390	434	111%	492	126%	434	111%
	HLT	50	50	100%	0	0%	n/a	n/a
	Sum	1360	1297	95%	1402	103%	1247	92%
Others		50	50	100%	26	51%	n/a	n/a
Total		**1,410**	**1,347**	**96%**	**1,428**	**101%**	**1247**	**92%**
Disk	Tier-0	30.3	30.3	100%	23.4	77%	15.6	51%
	Tier-1	60.5	54.7	90%	35.2	58%	55.0	91%
	Tier-2	11.6	7.9	68%	3.6	31%	7.3	63%
	Total	**102.5**	**92.9**	**91%**	**62.2**	**61%**	**77.8**	**76%**
Tape	Tier-0	91	91	100%	37.8	42%		
	Tier-1	157	134	85%	55.8	35%		
	Total	**248.3**	**225.0**	**91%**	**93.6**	**38%**		

Table A-0-1: Summary of the scrutinized, pledged, deployed resources in the 2023 WLCG year. The CPU resources used in the 2023 calendar year, and the occupancy of storage resources on 31/12/2023, are also reported. As CERN and RAL do not publish their tape buffer information, the Tier0 and Tier1 disk occupancies reported in the table below are lower than the real ones.