

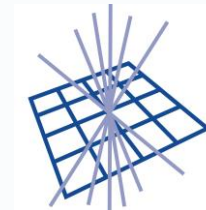
---

# LHCb status

---

GRIDPP 52

alexrg

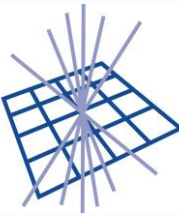


# Contents

---

In this talk I'll try to cover the following topics:

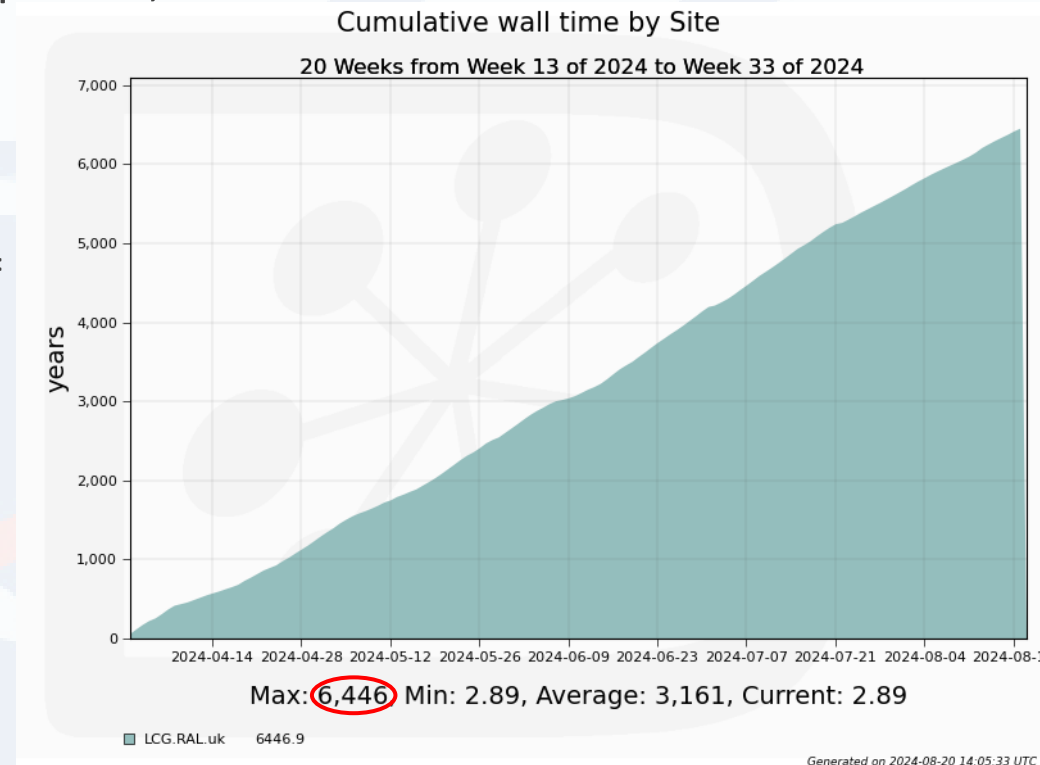
1. LHCb UK resource usage for the last half year.
2. UK sites problems and plans
  - Focusing mostly on RAL T1
3. LHCb news.



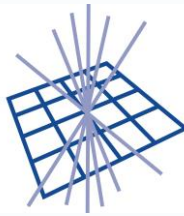
# LHCb Jobs at RAL T1

Nowadays it's a little tricky to calculate average HS23 consumption...

- DIRAC code for calculating normalization factor is being updated, so can not be trusted
- If we use static normalization factor from RAL, we have:
  - $6446 * 12.7 / 0.4 = 204660$  HS23
  - Blue = Wallclock years; purple = Normalization factor, green = years in the reporting period
  - Pledge is 180 kHS23

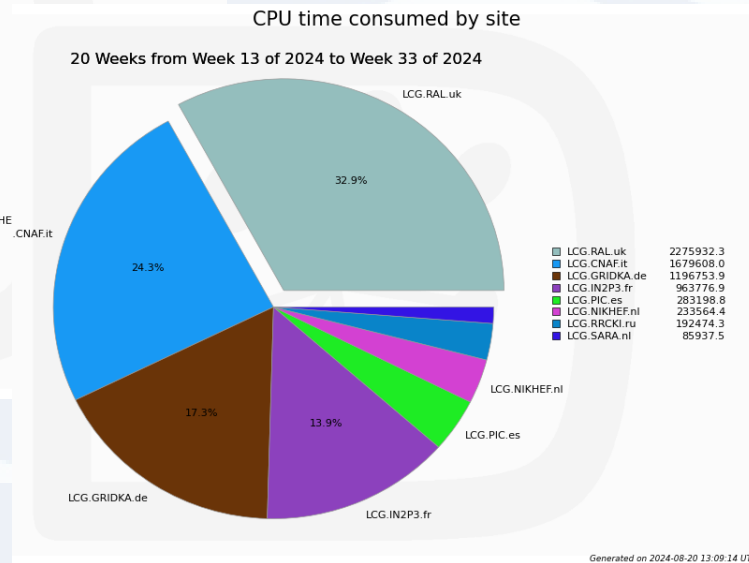
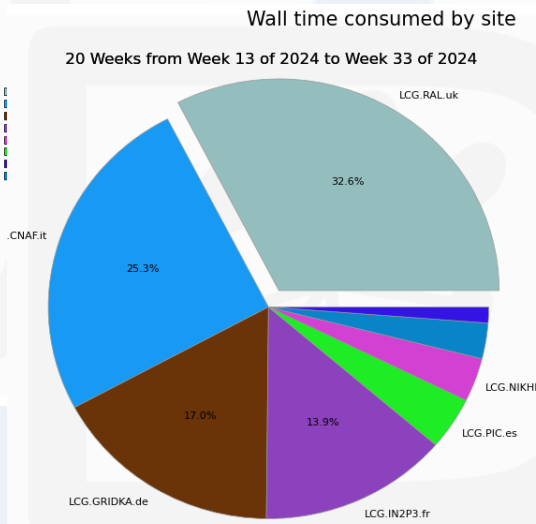
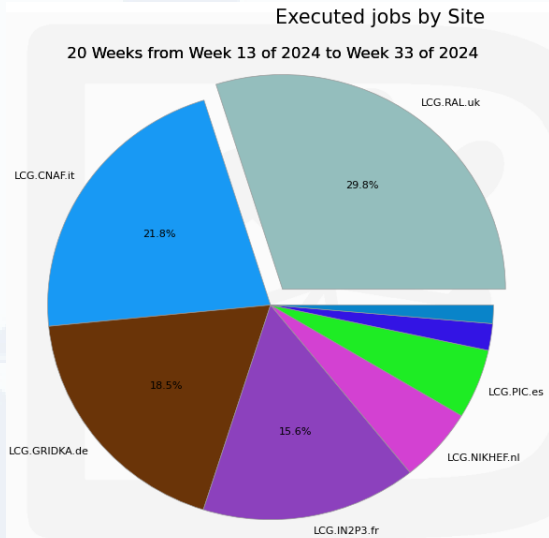


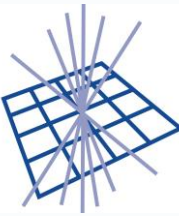
Generated on 2024-08-20 14:05:33 UTC



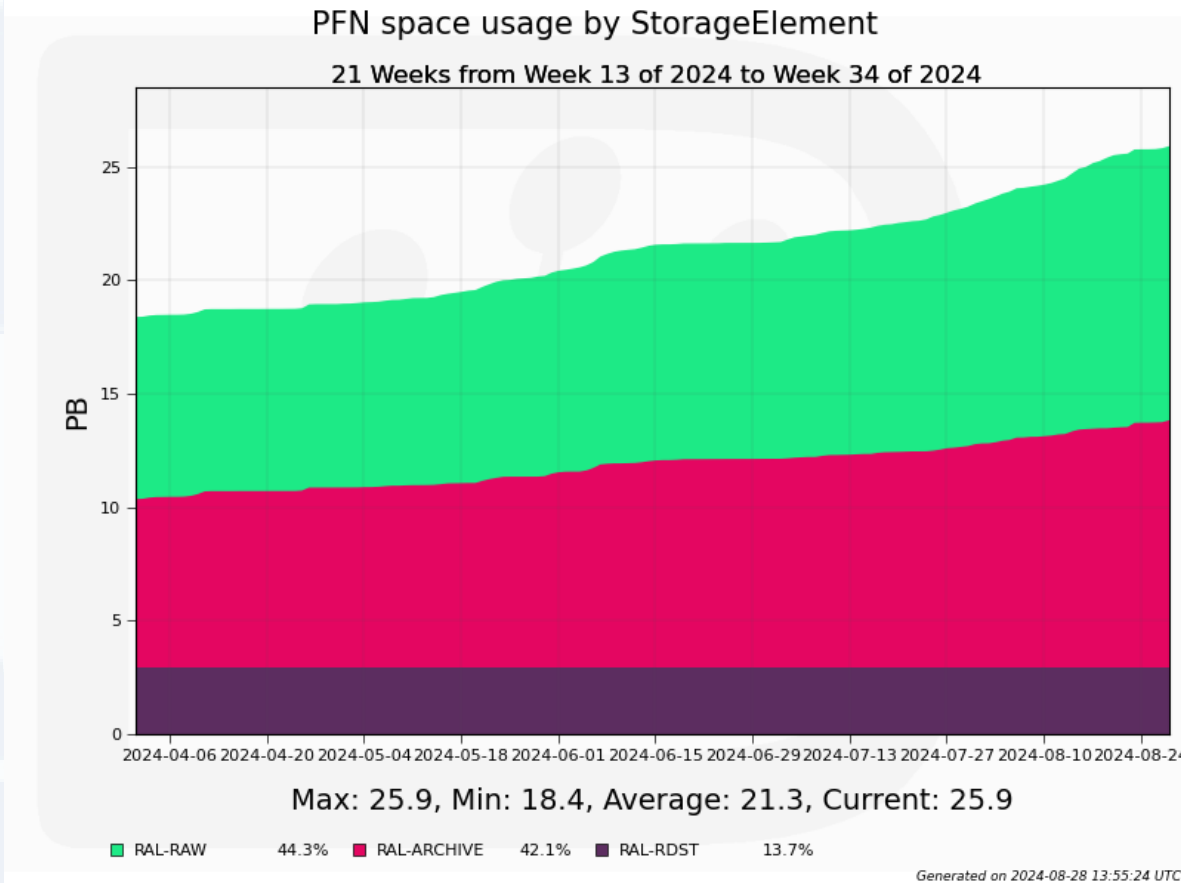
# Comparison

RAL has provided the biggest amount of CPU resources among all T1 sites, in terms of executed jobs (first plot), walltime (second), and cputime (third).

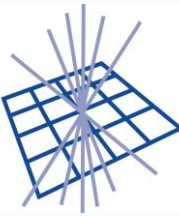




# Tape usage



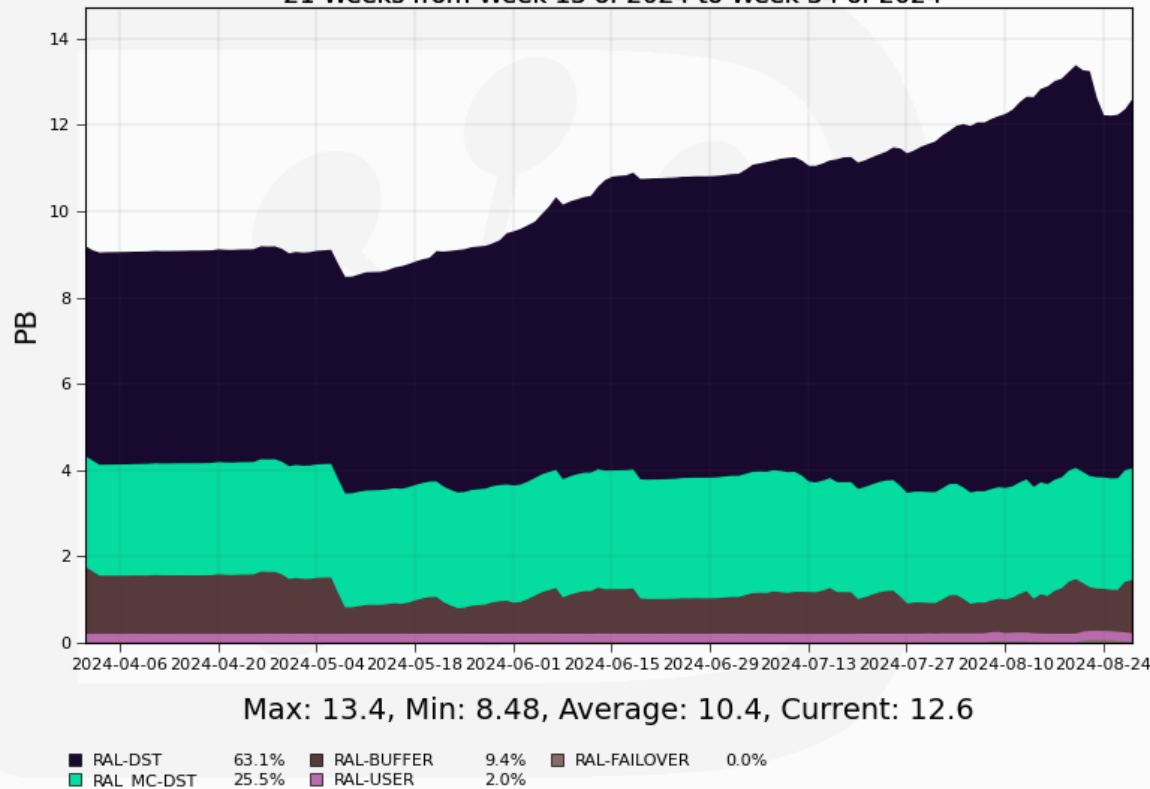
- Tape usage increased significantly
  - ~78% of the pledged space is used
    - Contrary to ~56% at the beginning of the Data Taking this year
  - Pledge for this FY is 33PB
  - Data Taking campaign is still ongoing!
    - Likely to use 7.2PB more if the rate remains the same



# Disk usage

PFN space usage by StorageElement

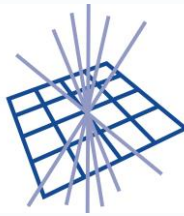
21 Weeks from Week 13 of 2024 to Week 34 of 2024



Max: 13.4, Min: 8.48, Average: 10.4, Current: 12.6

Generated on 2024-08-28 13:57:34 UTC

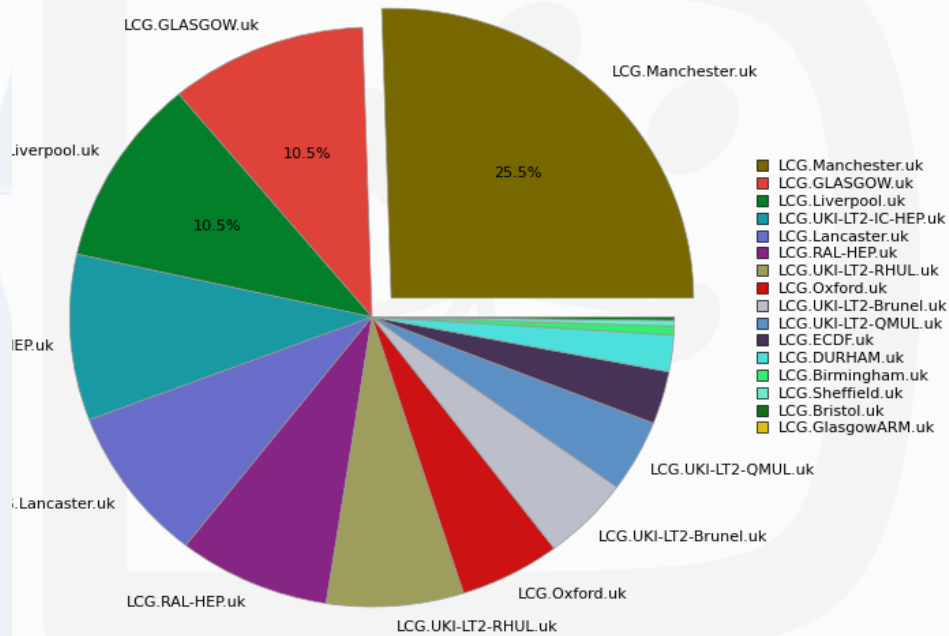
- Disk usage increased significantly
  - ~80% of the pledged space is used
  - Pledge for this FY is 15.7PB
  - Data Taking campaign, which is still ongoing!



# Tier-2 statistics

Wall time days used by Site

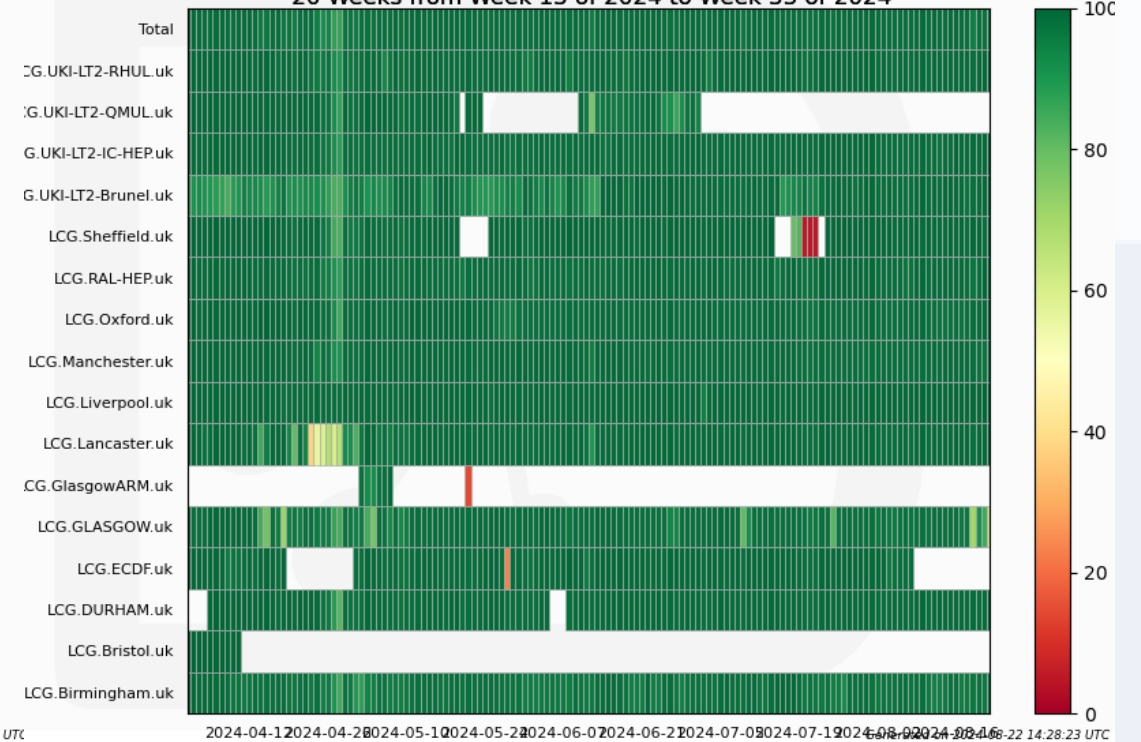
20 Weeks from Week 13 of 2024 to Week 33 of 2024



LCG.Manchester.uk	697567.3
LCG.GLASGOW.uk	288164.3
LCG.Liverpool.uk	286768.5
LCG.UKI-LT2-IC-HEP.uk	253351.6
LCG.Lancaster.uk	237170.5
LCG.RAL-HEP.uk	222493.9
LCG.UKI-LT2-RHUL.uk	200330.6
LCG.Oxford.uk	147257.6
LCG.UKI-LT2-Brunel.uk	129689.1
LCG.UKI-LT2-QMUL.uk	109682.2
LCG.ECDF.uk	80227.4
LCG.DURHAM.uk	55953.9
LCG.Birmingham.uk	13358.4
LCG.Sheffield.uk	8559.9
LCG.Bristol.uk	5085.8
LCG.GlasgowARM.uk	2.7

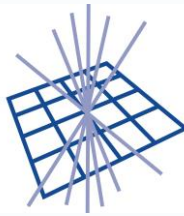
Job CPU efficiency by Site

20 Weeks from Week 13 of 2024 to Week 33 of 2024



Generated on 2024-08-22 14:22:33 UTC

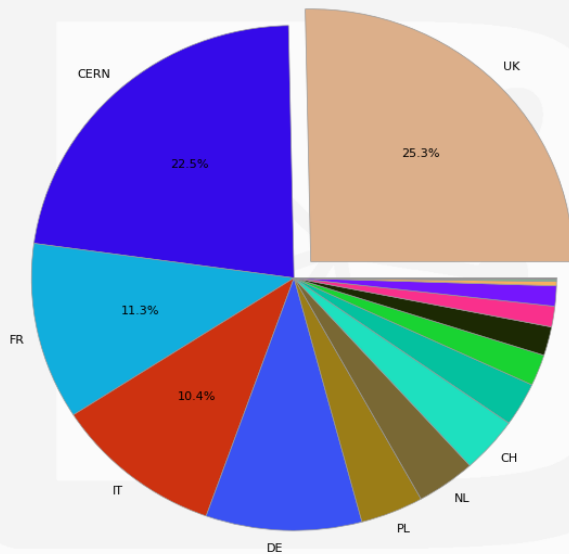
2024-04-12 2024-04-28 2024-05-10 2024-05-22 2024-06-07 2024-06-21 2024-07-09 2024-07-16 2024-08-01 2024-08-16 2024-08-22 14:28:23 UTC



# Tier-2 statistics

Total Number of Jobs by Country

20 Weeks from Week 13 of 2024 to Week 33 of 2024

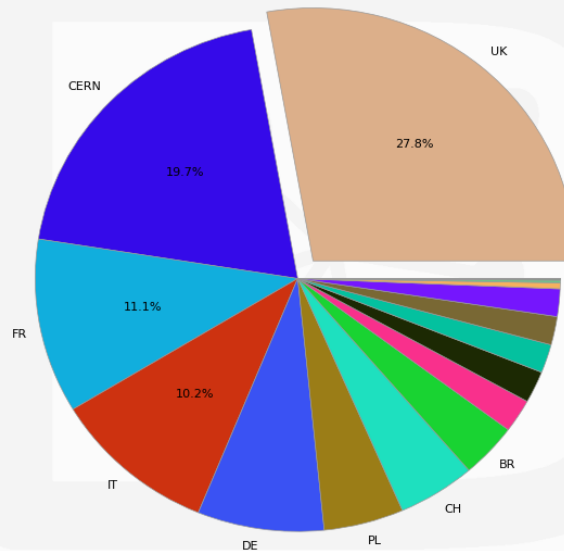


UK provides more computing resources than CERN!

- According to the pledges, though
- Wall/CPU time fraction relation is higher than the pledge

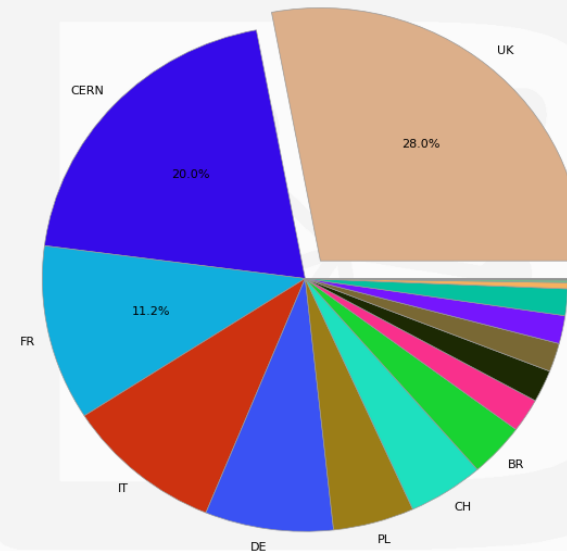
Wall time days used by Country

20 Weeks from Week 13 of 2024 to Week 33 of 2024



CPU days used by Country

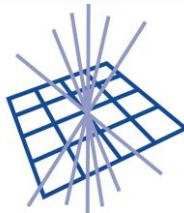
20 Weeks from Week 13 of 2024 to Week 33 of 2024



UK	4997314.7
CERN	3565311.4
FR	2000405.8
IT	1734872.3
DE	1409328.3
PL	891171.2
CH	832459.8
BR	615375.6
RU	385486.2
CN	366390.5
NL	326685.8
US	320445.3
ES	303468.5
RO	70750.4
AU	26204.8
IL	9841.1
CR	9013.5
MULTIPLE	129.8
ANY	46.8

Generated on 2024-08-22 14:31:57 UTC





# Problems

At the time of writing, there are 5 opened LHCb tickets against UK sites:

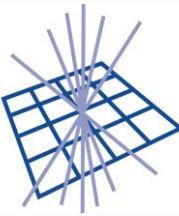
**5 of 5 Tickets**

Ticket-ID	Type	VO	Site	Priority	Resp. Unit	Status	Last Update	Subject	Scope
<a href="#">167910</a>	Team	lhcb	UKI-LT2-Brunel	very urgent	NGI_UK	in progress	2024-08-20	Jobs Failed at UKI-LT2-Brunel	WLCG
<a href="#">167852</a>	Team	lhcb	UKI-SOUTHGRID-OX-HEP	very urgent	NGI_UK	in progress	2024-08-12	Pilots failing at Oxford	WLCG
<a href="#">167682</a>	Team	lhcb	UKI-SCOTGRID-GLASGOW	very urgent	NGI_UK	in progress	2024-08-21	All FTS transfers to Glasgow are failing	WLCG
<a href="#">167007</a>	Team	lhcb	UKI-SCOTGRID-ECDF	very urgent	NGI_UK	in progress	2024-08-13	Pilots Failed UKI-SCOTGRID-ECDF	WLCG
<a href="#">163853</a>	Team	lhcb	UKI-NORTHGRID-SHEF-HEP	urgent	NGI_UK	on hold	2024-07-24	Failed jobs at LCG.Sheffield.uk	WLCG

Most of them are awaiting closure, the last one (for Sheffield) is awaiting actions from LHCb.

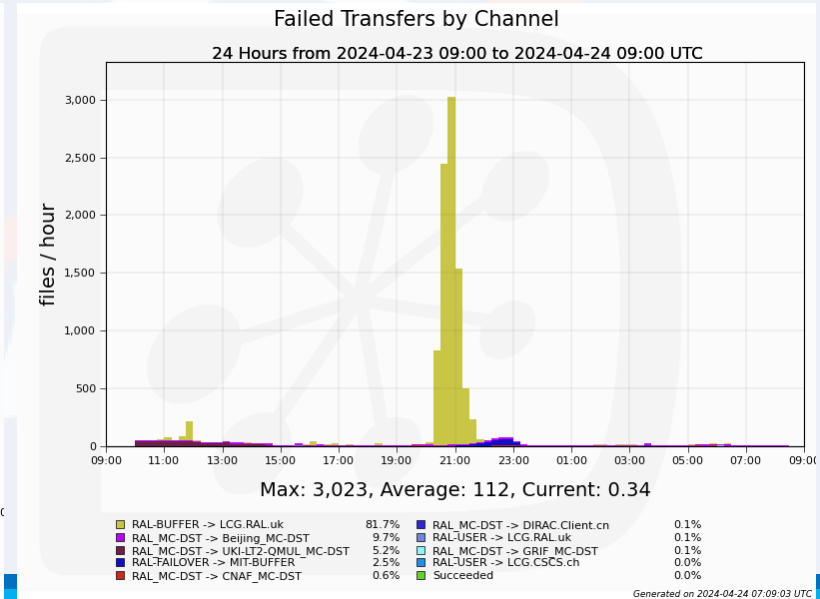
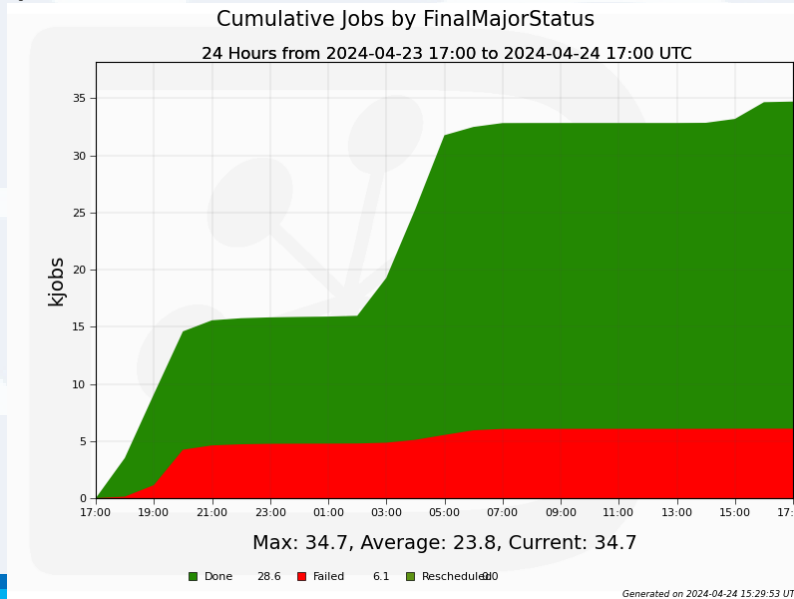
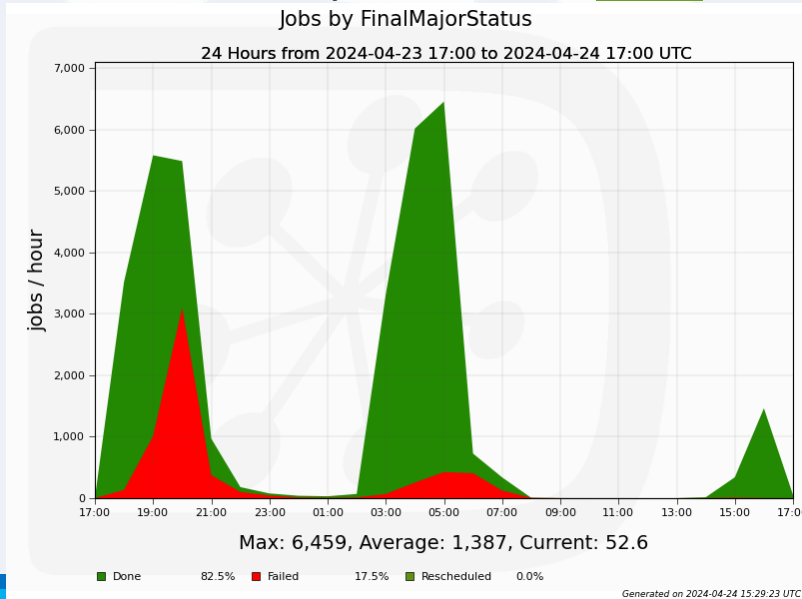
At RAL there is still one (or “half”?) long-lasting issue for LHCb:

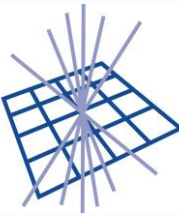
1. Direct access issues ([ticket](#), closed, [new ticket](#), closed).



# Direct access issues

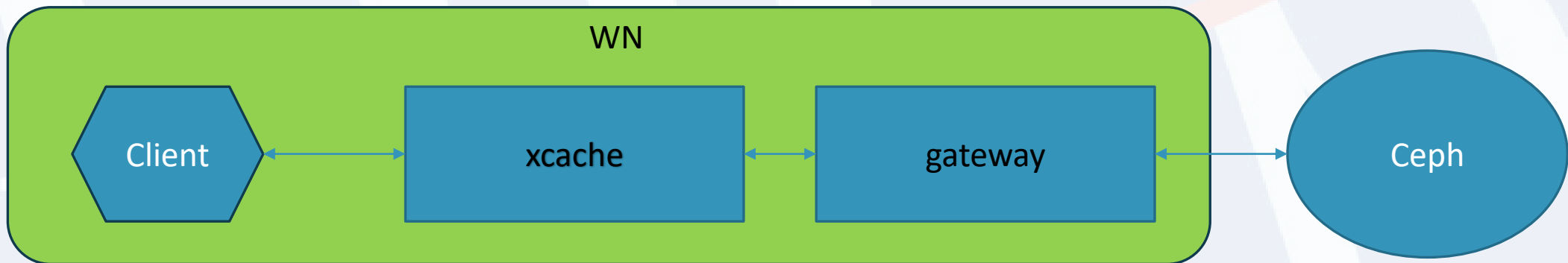
- Some LHCb jobs (namely user and WGProduction) do not download data, but access it directly from the storage using xrootd's (vector) read requests
- Vector read requests were problematic for ECHO for a long time, until ~Aug 2023
- This year the number of WGProduction jobs at RAL increased, and significant failures started to appear occasionally (when large WGProds are submitted), causing also download failures (see below)
- In July 2024 a new [ticket](#) was opened

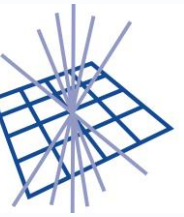




# Direct access issues: memory

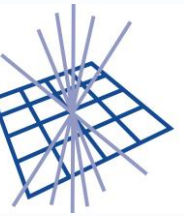
- Most of the errors were happening on the 2021 Generation of WNs
  - WNs with 256 slots and SATA SSDs
- The most popular error was “Cannot allocate memory”, reported by xrootd proxy
  - It was confirmed that Proxy’s memory consumption can be close to its limit
- Failed downloads were caused by Ceph being overload with IOps
  - When proxy runs out of memory, it forwards read requests directly to the gateway
  - Ceph does not like when it is asked to do many small reads





# Direct access issues: memory

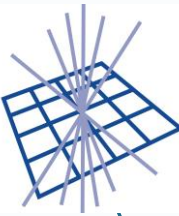
- Turned out that proxy's memory limit(s) was(were) low (8GiB) and independent on the number of cores
  - Both container limit and xrootd limit
- The increase of the limit reduced ECHO IOps, but did not help to solve (Vector) read errors
- Additional changes were applied to reduce proxy's memory consumption
  - Prefetching was turned off
    - The change was planned long time ago for another reason, but could have helped to deal with memory consumption as well
  - LHCb jobs were moved from 2021 Gen to 2022
    - 2022 Gen has NVMe SSDs, allowing faster writes (and therefore faster memory release)
- All was in vain...
  - Errors still happened occasionally



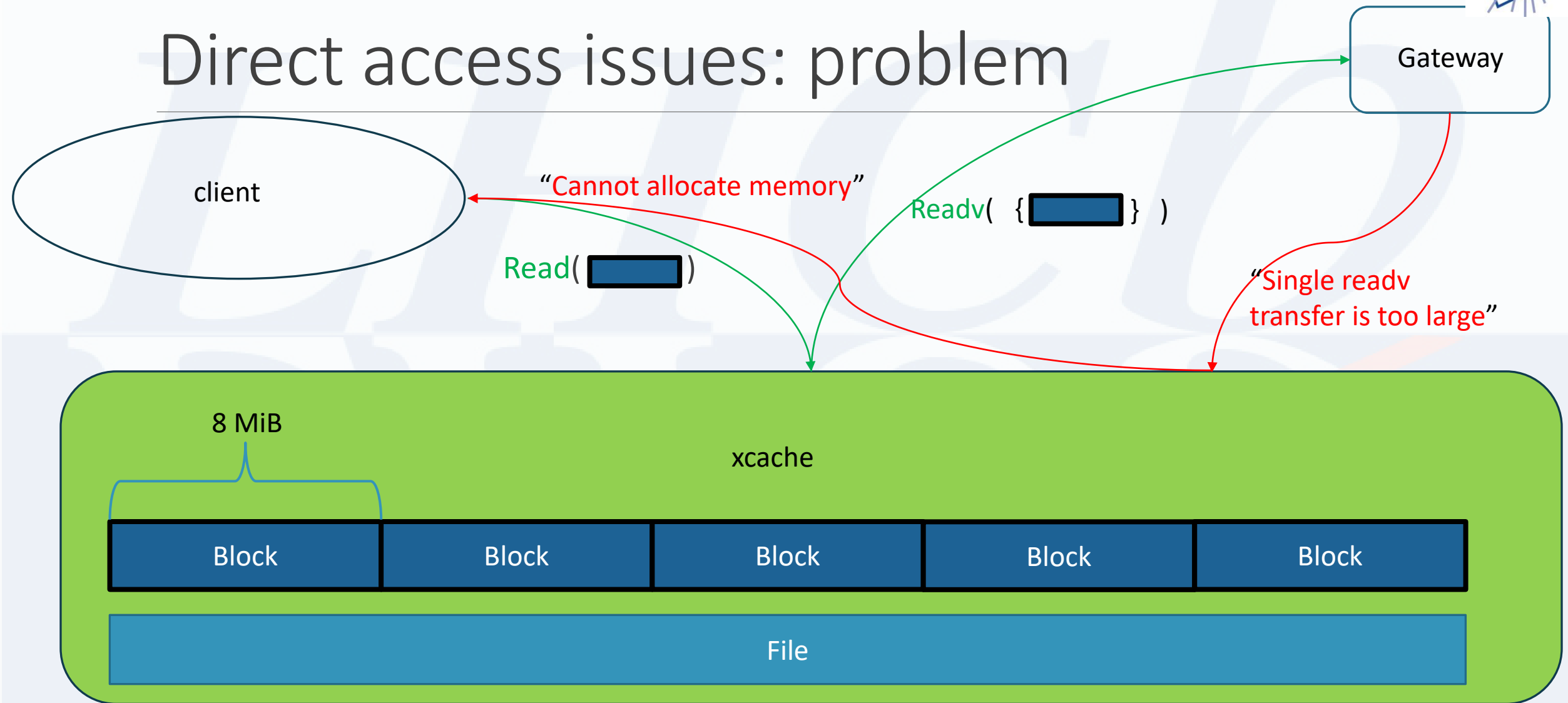
# Direct access issues: problem

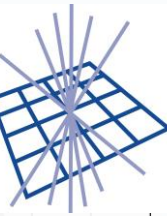
---

- Luckily, it turned out that the error can be reproduced manually
  - To do so one has to run many “jobs” doing ReadV requests + a job that is doing ordinary file downloads
- Turned out that the error was caused by a bug in xrootd and is not a simple memory excess
  - When proxy runs out of memory, it forwards all incoming requests to the gateway
  - Read requests are converted to ReadV, where read vector contains only one element
    - Conversion is very simple: Read coordinates are put into read vector, no other changes are made
  - In XrootD there is a limit on max chunk size in read vector in ReadV; for Read requests there is no limit
    - The limit is lower than proxy’s block size and XRootD’s buffer size (defines internal read lengths)
  - If a Read request is big enough, it can exceed the limit when converted to ReadV
    - When this happens, gateway sends an error `“Single readv transfer is too large”`
    - Due to XrootD creativeness, it becomes `“Cannot allocate memory”` when forwarded to the client



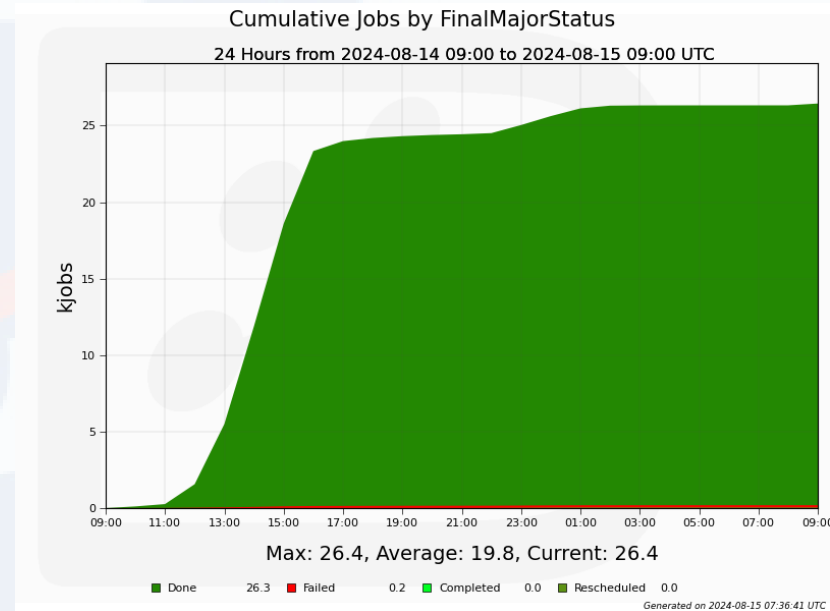
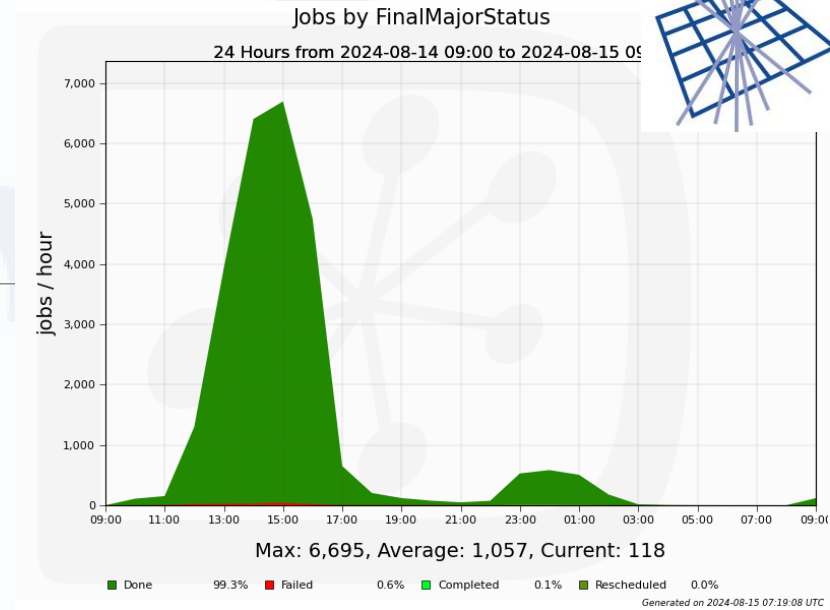
# Direct access issues: problem

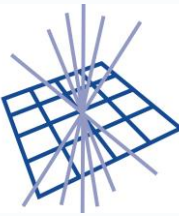




# Direct access issues: solution

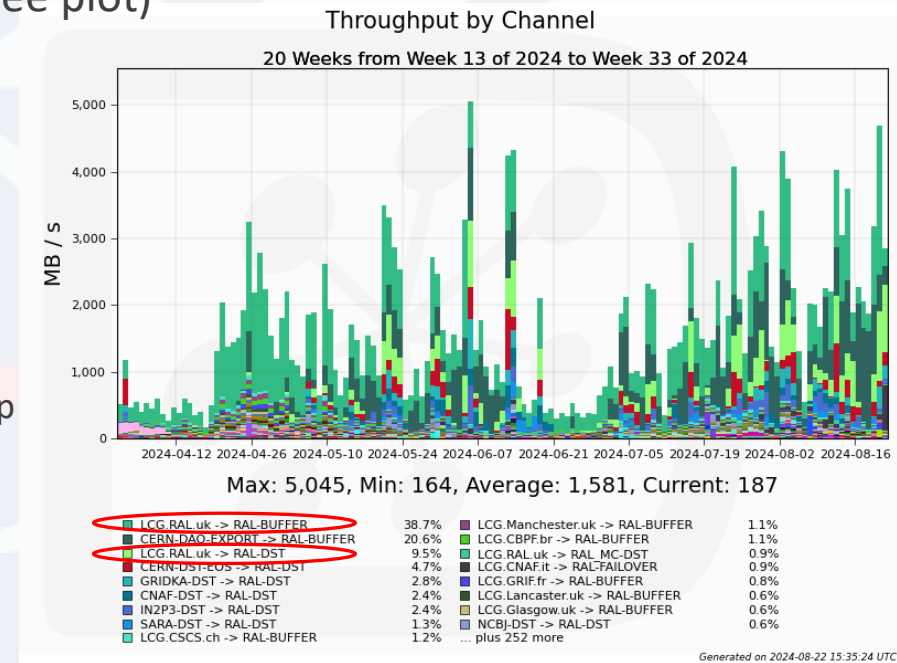
- The issue can be mitigated by turning Paged Read requests on
  - In this case file downloads will use smaller block size
  - Direct access requests still can trigger the issue though
- A github [issue](#) is opened to fix the problem properly
- Other issues were affecting Vector read performance as well
  - Genuine lack of memory on WN gateways
    - Fixed by additional memory allocation
  - Daily restarts of xrootd services on WNs
    - Jobs became less tolerant to these restarts after prefetch had been turned on
    - Fixed by redesigning the restart script
- ~~Once the github issue is resolved, we can call the problem solved~~
  - Github issue is resolved, the fix should be present in v5.7.1



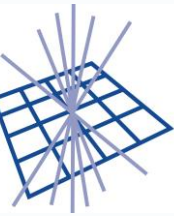


# Further WN XRootD improvements

- Multiple times we have seen external gateways using their full network throughput
  - E.g. during DC24
- These gateways are used for WN uploads as well, since proxies on WNs are read-only
  - For LHCb these WN uploads are the biggest throughput consumer (see plot)
- Can we make local WN gateways writable?
  - In principle, it is possible to patch proxy so that it forwards all write requests to the gateway
    - However, there are a few things to tune
      - Authentication: currently gateways can not authenticate clients via X509/tokens
      - Writing protocols: currently LHCb writes only via https
        - Since xrootd writes can cause file loss due to retries and our multi-gateway setup
          - Potentially can be overcome by turning retries off or setting up https on the WN gateways
  - Remove proxy completely?
    - Increases IOps on ECHO significantly, so should be tested with care



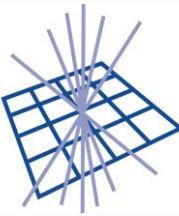




# LHCb News: future plans

- There is a significant increase in computing resources requests for the next FY [1]
  - 62% increase for CPU, 57% for Disk, 46% for Tape (relative to this FY)
  - During the LS3 there expected to be no increase at all
  - After this increase further resource requirement should stay within “Flat Cache” budget model
- Possible extension of the Run 3 introduces some uncertainty..
- Two new Tier-1 centers (Beijing and NCBJ) will be providing resources as well
- Network capacity seems to be already good enough
  - Should stay the same order of Magnitude as during DC24 until the end of run 4
- ARM usage is not ready for production yet

[1]<https://indico.cern.ch/event/1389300/contributions/5840870/attachments/2833421/4950927/CERN-RRB-2024-012.pdf>

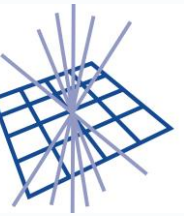


# LHCb News: ETF tests

- New set of tests is used in production since May ([see here](#) and [here](#))
  - Finally, includes storage tests!
- There are still a few issues, namely
  - ARC client bug is causing intermittent failures for CE tests
  - Container that is used for running test infrastructure is still based on CentOS7 image
    - We can not do much, since base container is maintained separately
    - All Vos are affected
- Future development plans:
  - Fix issues!
  - Add token-based SE tests (for me)
    - Currently tokens are only used for job submissions to HTCondor CEs

Production

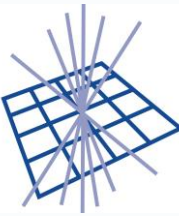
[LCG.RAL.uk] [ARC-CE] arc-ce-test02.gridpp.rl.ac.uk: NODATA (0.02%) OK (78.10%) CRITICAL (10.83%) UNKNOWN (11.04%)
[LCG.RAL.uk] [ARC-CE] arc-ce01.gridpp.rl.ac.uk: NODATA (0.02%) OK (78.59%) CRITICAL (10.28%) UNKNOWN (11.11%)
[LCG.RAL.uk] [ARC-CE] arc-ce02.gridpp.rl.ac.uk: NODATA (0.02%) OK (81.58%) CRITICAL (7.36%) UNKNOWN (11.04%)
[LCG.RAL.uk] [ARC-CE] arc-ce03.gridpp.rl.ac.uk: NODATA (0.02%) OK (78.26%) CRITICAL (10.60%) UNKNOWN (11.11%)
[LCG.RAL.uk] [ARC-CE] arc-ce04.gridpp.rl.ac.uk: NODATA (0.02%) OK (83.31%) CRITICAL (5.63%) UNKNOWN (11.04%)
[LCG.RAL.uk] [ARC-CE] arc-ce05.gridpp.rl.ac.uk: NODATA (0.02%) CRITICAL (9.65%) OK (79.15%) UNKNOWN (11.18%)
[LCG.FLORUK] [XROOTD] antares.stfc.ac.uk: NODATA (0.02%) CRITICAL (2.99%) DOWNTIME (0.83%) OK (96.16%)
[LCG.RAL.uk] [XROOTD] xrootd.echo.stfc.ac.uk: NODATA (0.02%) OK (99.98%)



# LHCb News: UK DC

---

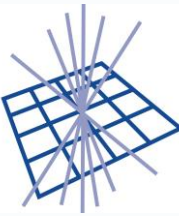
- After the Data Taking, Reprocessing Campaign starts
  - Lasts almost until the next year's DT starts
- UK-wide DC is likely to overlap with LHCb's reprocessing campaign
  - Tape usage is undesirable
    - Since it will interfere with reprocessing staging
  - Disk usage is OK as long as
    - We have enough space to accommodate test and production data
    - Testing does not interfere with production activities



# Conclusion

---

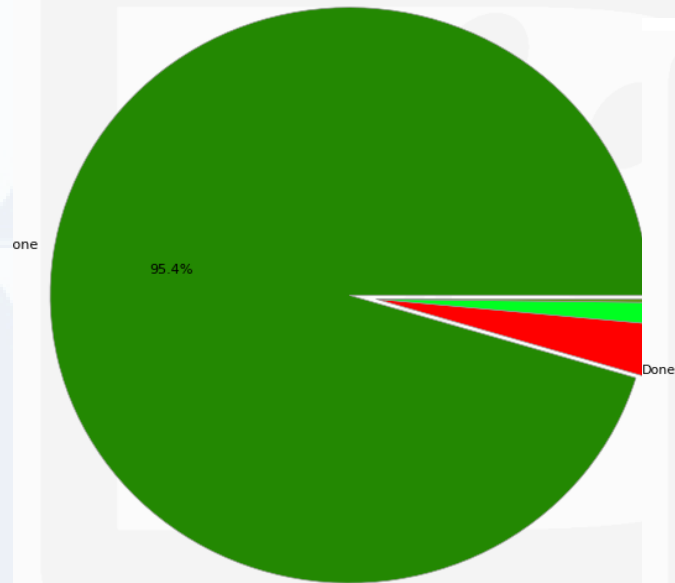
- UK provided a lot of resources to LHCb during the last 5 months
  - Around 30%
- Relatively smooth operations
- Long lasting issues are (almost completely) resolved
- New developments are ongoing



# RAL T1 Job success rate (backup)

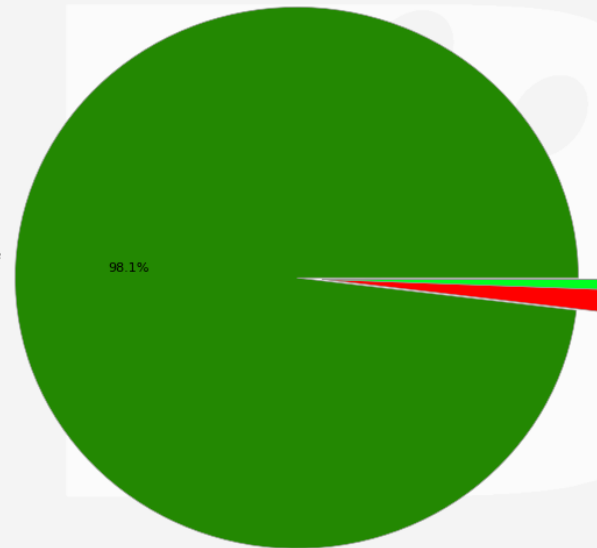
RAL jobs by Status

20 Weeks from Week 13 of 2024 to Week 33 of 2024



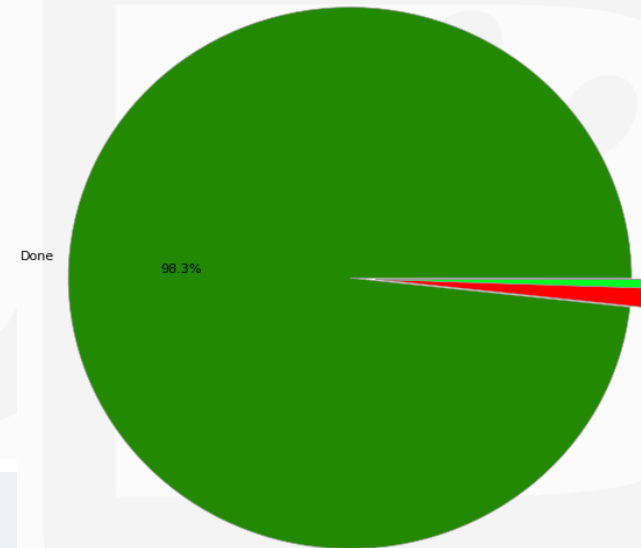
RAL wall time consumed by status

20 Weeks from Week 13 of 2024 to Week 33 of 2024



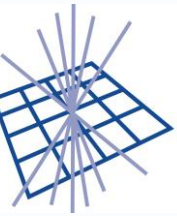
RAL CPU time consumed by status

20 Weeks from Week 13 of 2024 to Week 33 of 2024



Done	2236871.0
Failed	27024.3
Completed	12034.7
Rescheduled	2.2

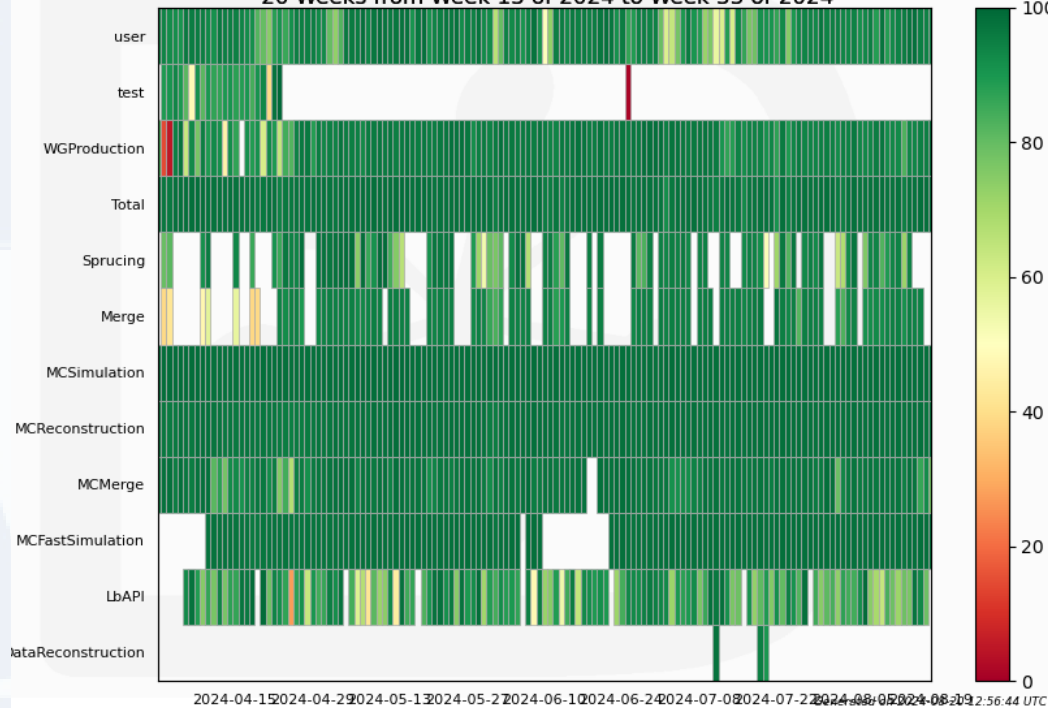
Generated on 2024-08-20 13:03:15 UTC



# RAL T1 CPU efficiency (backup)

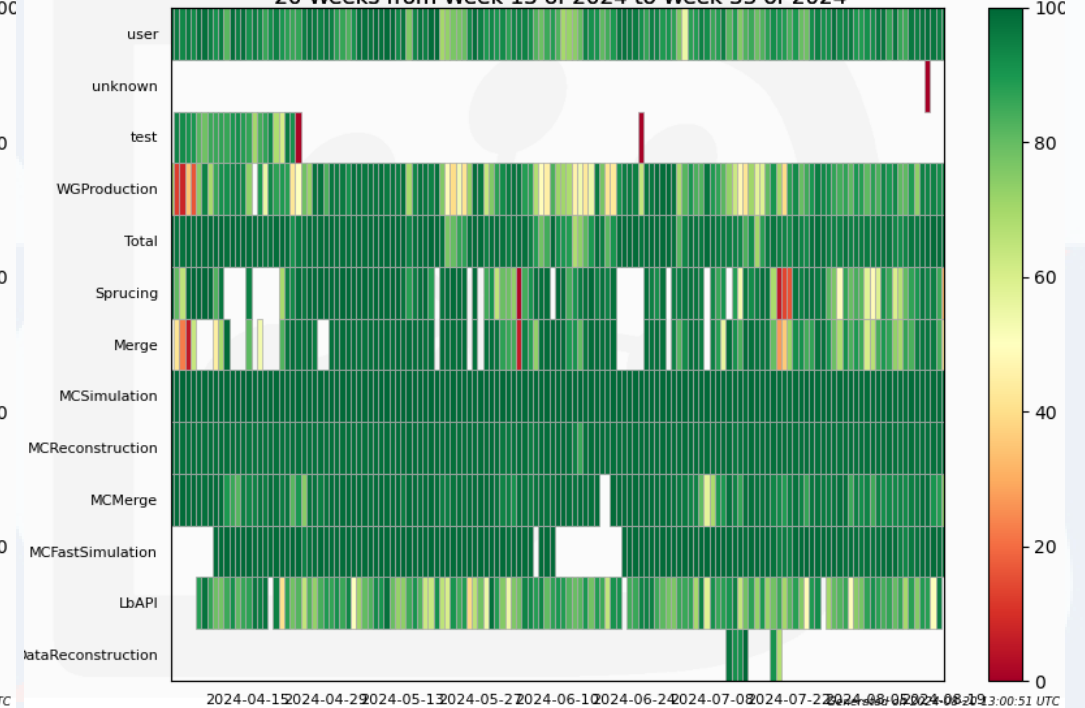
CPU efficiency (RAL)

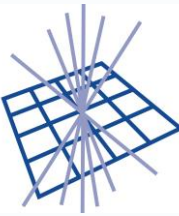
20 Weeks from Week 13 of 2024 to Week 33 of 2024



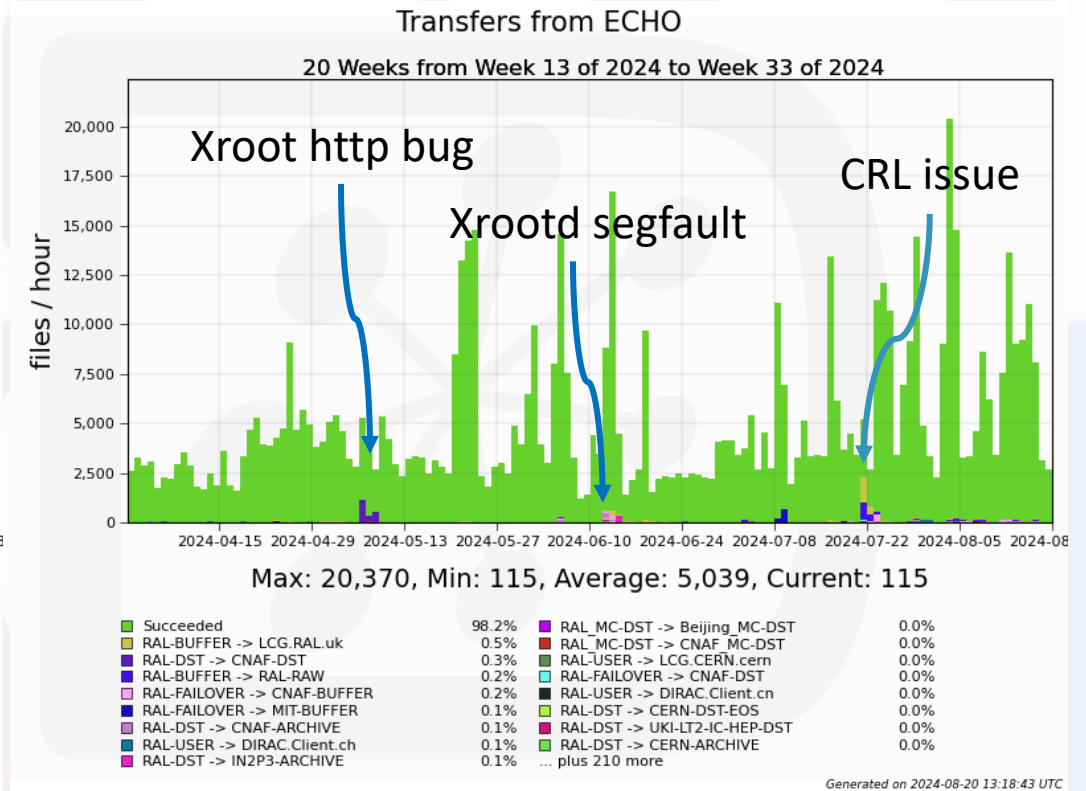
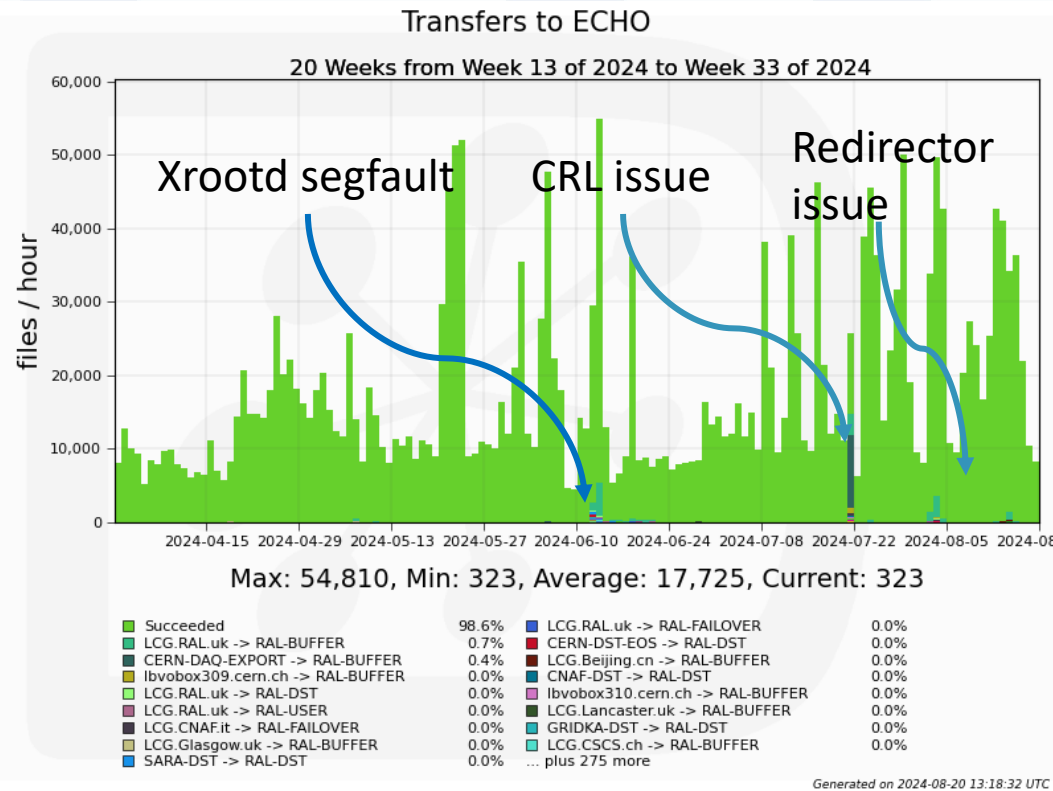
CPU efficiency (T1s except RAL)

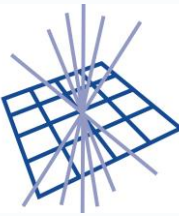
20 Weeks from Week 13 of 2024 to Week 33 of 2024





# Transfers to ECHO (backup)

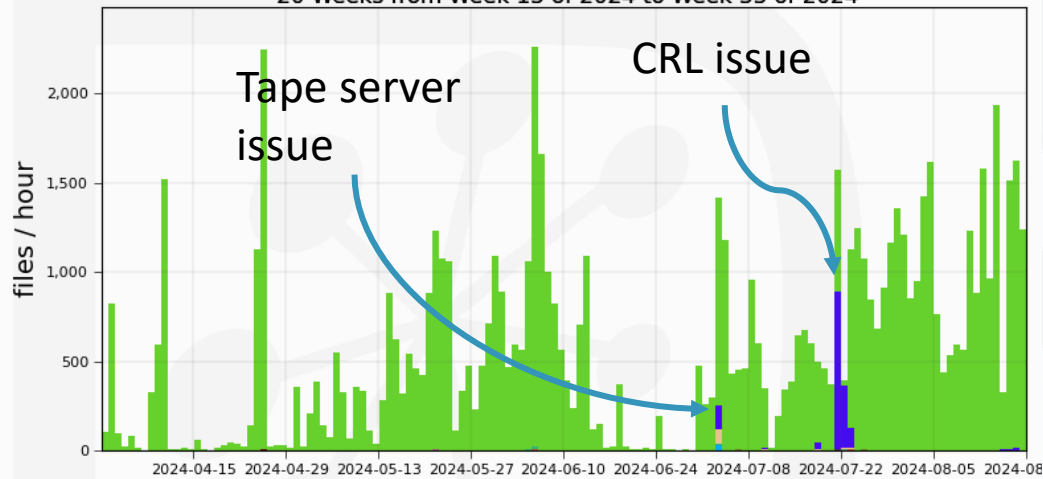




# Transfers to Antares (backup)

Transfers to Antares

20 Weeks from Week 13 of 2024 to Week 33 of 2024



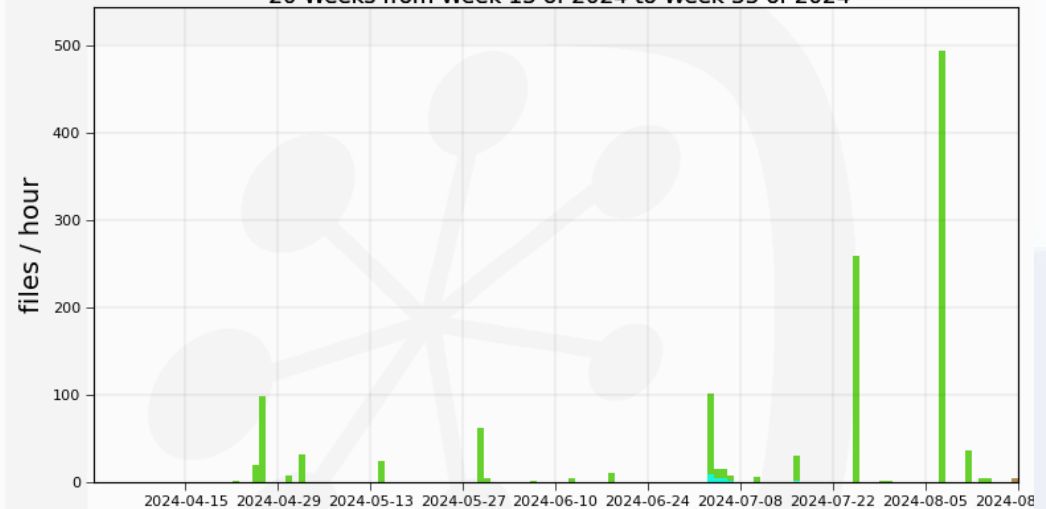
Max: 2,258, Average: 544, Current: 115

■ Succeeded	97.7%	■ GRIF-DST -> RAL-ARCHIVE	0.0%
■ RAL-BUFFER -> RAL-RAW	2.0%	■ CERN_MC-DST-EOS -> RAL-ARCHIVE	0.0%
■ CERN-DAQ-EXPORT -> RAL-RAW	0.2%	■ IN2P3-DST -> RAL-ARCHIVE	0.0%
■ CERN-DST-EOS -> RAL-ARCHIVE	0.0%	■ GRIDKA_MC-DST -> RAL-ARCHIVE	0.0%
■ SARA-DST -> RAL-ARCHIVE	0.0%	■ RAL-DST -> RAL-ARCHIVE	0.0%
■ CNAF-DST -> RAL-ARCHIVE	0.0%	■ IN2P3_MC-DST -> RAL-ARCHIVE	0.0%
■ GRIDKA-DST -> RAL-ARCHIVE	0.0%	■ CNAF_MC-DST -> RAL-ARCHIVE	0.0%
■ RAL_MC-DST -> RAL-ARCHIVE	0.0%	■ PIC-DST -> RAL-ARCHIVE	0.0%
■ NCB-DST -> RAL-ARCHIVE	0.0%	... plus 34 more	

Generated on 2024-08-20 13:18:53 UTC

Transfers from Antares

20 Weeks from Week 13 of 2024 to Week 33 of 2024



Max: 494, Average: 8.97, Current: 4.83

■ Succeeded	97.7%	■ RAL-ARCHIVE -> CERN_MC-DST-EOS	0.0%
■ RAL-RAW -> RAL-BUFFER	1.5%	■ RAL-ARCHIVE -> RAL-ARCHIVE	0.0%
■ RAL-ARCHIVE -> DIRAC.Client.cn	0.4%	■ RAL-RDST -> RAL-BUFFER	0.0%
■ RAL-ARCHIVE -> DIRAC.Client.ch	0.3%	■ RAL-ARCHIVE -> CSCS-DST	0.0%
■ RAL-RAW -> DIRAC.Client.ch	0.0%	■ RAL-ARCHIVE -> GRIF-DST	0.0%
■ RAL-ARCHIVE -> DIRAC.Jenkins.ch	0.0%	■ RAL-ARCHIVE -> RAL-HEP-DST	0.0%
■ RAL-ARCHIVE -> DIRAC.Client.uk	0.0%		

Generated on 2024-08-20 13:18:58 UTC