January, 17th 2025

Report LHCb

Lucio Anderlini
INFN Firenze

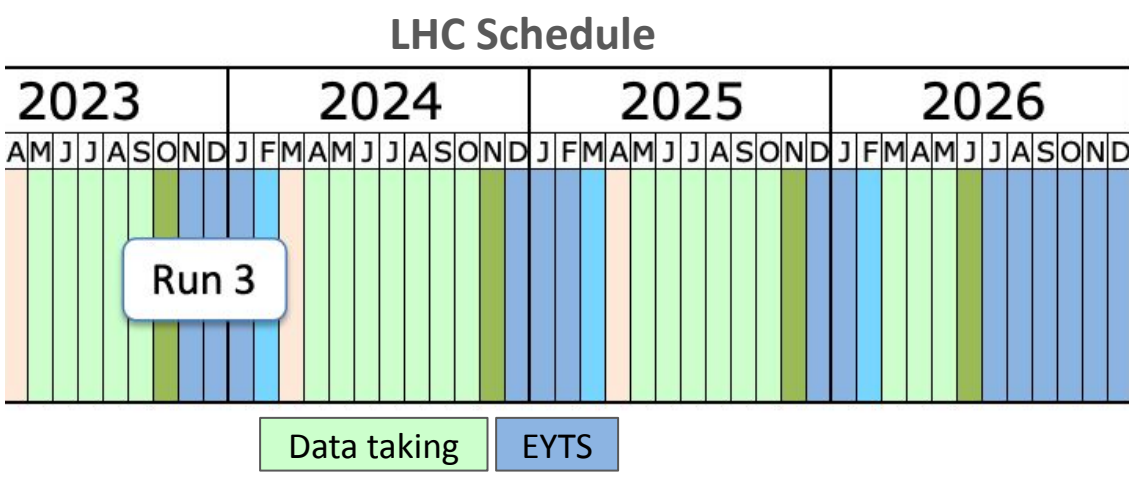
Rosa Petrini
INFN Firenze

LHCb Status

Since early December, LHCb is in its "***Reprocessing phase***":

- read data from tape (@CNAF: 4 PB)
- apply bug-fixes and new features to reconstruction code to 2024 data
- store re-reconstructed data on tape

Reprocessing must be 100% completed by mid April.



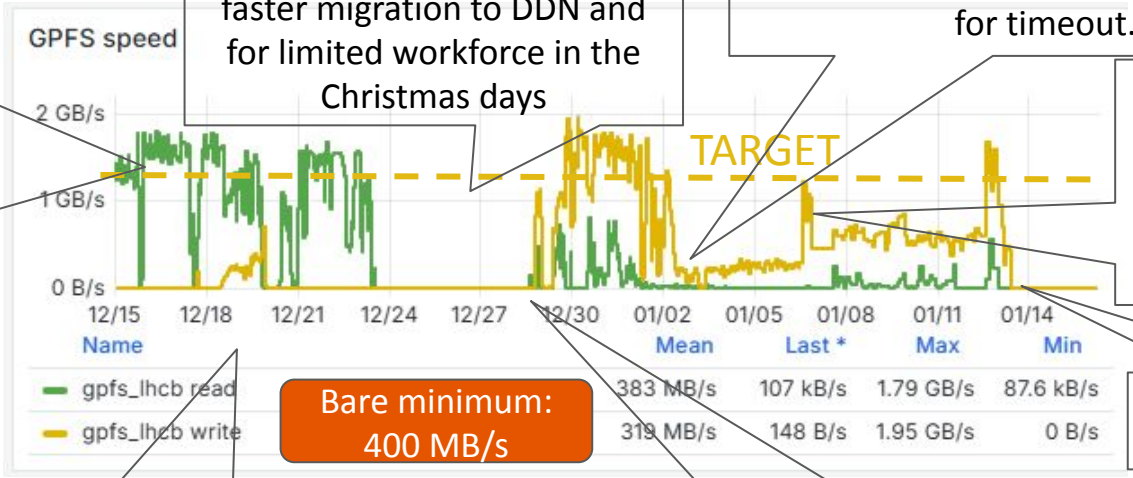
From mid-December to mid-April: ~10^7 s
4×10^6 GB / 10^7 s = **0.4 GB/s**
On average, with no contingency and 100% availability and operations.
DC-target was set to **1.3 GB/s** for contingency, operations, downtimes… and because a larger share was assigned to CNAF.

A passionate Christmas break

Backlog recovery of the ***processing phase*** *due to* downtime for tape library relocation (1 week) + fixing instabilities

LHCb keeps the ban to allow faster migration to DDN and for limited workforce in the Christmas days

After 3 days, **garbage collector drops the data from buffer before FTS could copy them in the disk file system.** FTS triggers transparent BOL and fails for timeout.



Bare minimum: 400 MB/s

New collapse of nginx (Tape API unreachable), due to network problems in the DDN cluster.

Downtime for library relocation

The reprocessing campaign starts, but CNAF gets banned almost immediately to a very high rate of TPC errors. **Probably** due to the migration to DDN.

CNAF asks removing the ban and restarting with staging as DDN migration is slow for external reasons.
BOL requests are submitted.
Some nginx error promptly fixed by Andrea

FTS: our thermometer for operations burden

Event during mid-throughput days, the success rate of TPC reported by FTS is very low.

Managing large failure rates increases the needs for human intervention making operations more stressful.

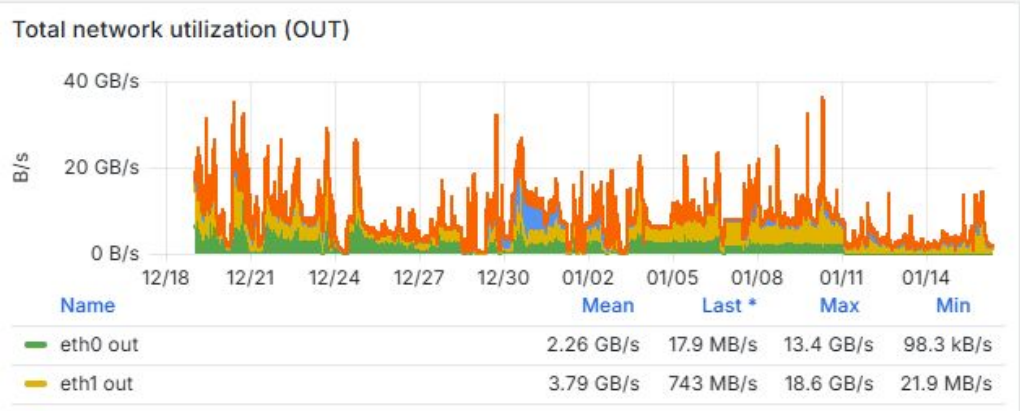
On the 30th and 31st/12 the system was operational.

Fast nVME storage

During the reprocessing we have seen the nVME hot storage drives in action for the first time.

The integrated experience (2 days) is still not sufficient to draw solid conclusions, but it seems capable of **mitigating high iops from the farm for spiky operations** (50 TB/day of small temporary files).

Collecting **more metrics** may lead to contribution relevant to other LHCb T1s.



Corrupted files

POSIX access to data files is known to be much less error prone wrt. https.

Unfortunately, when accessing via POSIX, adler checksum is not computed and corruptions are not identified in standard reading (most critically, transfers to tape).

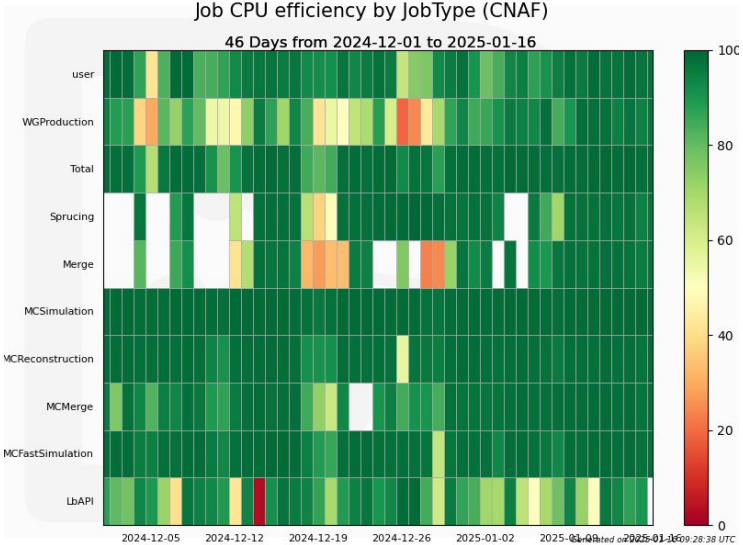
This caused some troubles with 1696 files (461 GB) as tracked here: GGUS#169555

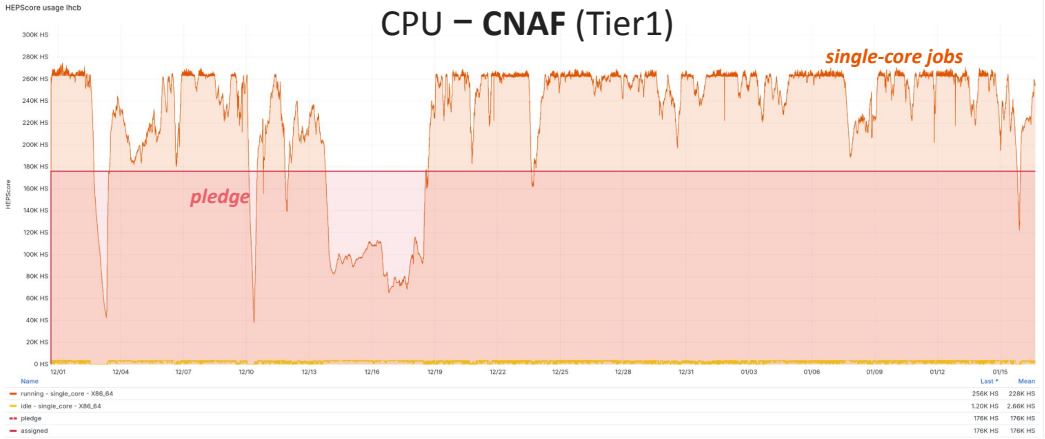
When we transitioned from https to posix we had 4 webdav servers and no hot storage partition.

At the end of the reprocessing campaign (April), with 8 servers and the optimization work done on WebDAV, we should probably schedule some exercise to move back from posix to https…

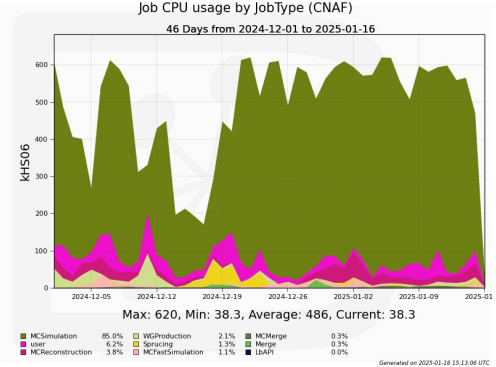
CPU usage

The use of CPU resources
is at **nominal values**

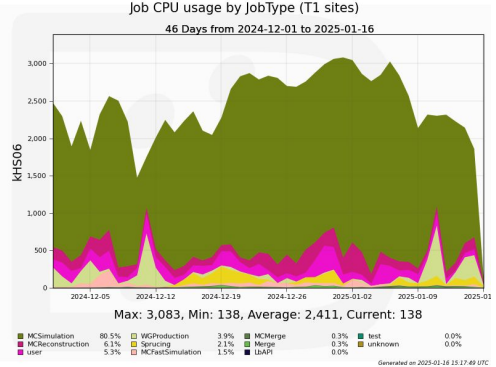




from DIRAC: **CNAF** (Tier1 + Tier2)



from DIRAC: **all** Tier1 sites



Backup

History of corrupted files

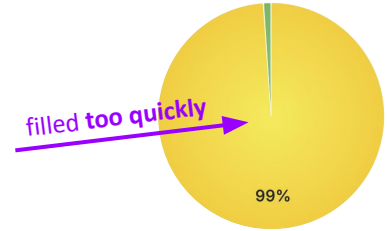
- **11-13 Sept 2024** Tier-1 experienced some **network issues** (GOCDB:35910) due to a hardware problem with a core switch Arista at Technopole
- **25 Sept 2024** A network intervention was needed to restore the nominal conditions after the problem raised on September 11th → **the intervention caused network instability**
- **1 Oct 2024** **Checksum inconsistency** between source file and file at CNAF found by Chris (ELOG:39126)
- **2 Oct 2024** Comparing source file with the one at CNAF, it seems that some **corruption actions** occurred at CNAF at certain point of the data transfer, **probably due to the network issues**
- **16 Oct 2024** The storage team performed an **intense research campaign** checking more than 80k files on disk and 3.5k files on tape and identified 7 corrupted files (GGUS:168495)

Hot-storage region

The scope of the ***hot*-region** is to (temporary) store files produced by Sprucing jobs which, processed by Merge jobs, are combined into a single merged file, then deleting all the input files used.

To pursue this goal, GPFS has been configured with a dedicated **placement policy**:

- `/gpfs_lhcb/disk/lhcb/**/*.dst` → NVMe (but this is the whole disk)
- `/gpfs_lhcb/disk/lhcb/buffer/**/*.dst` → NVMe (correct path)



`buffer/` isn't a fileset hence **no policy** can be set in GPFS