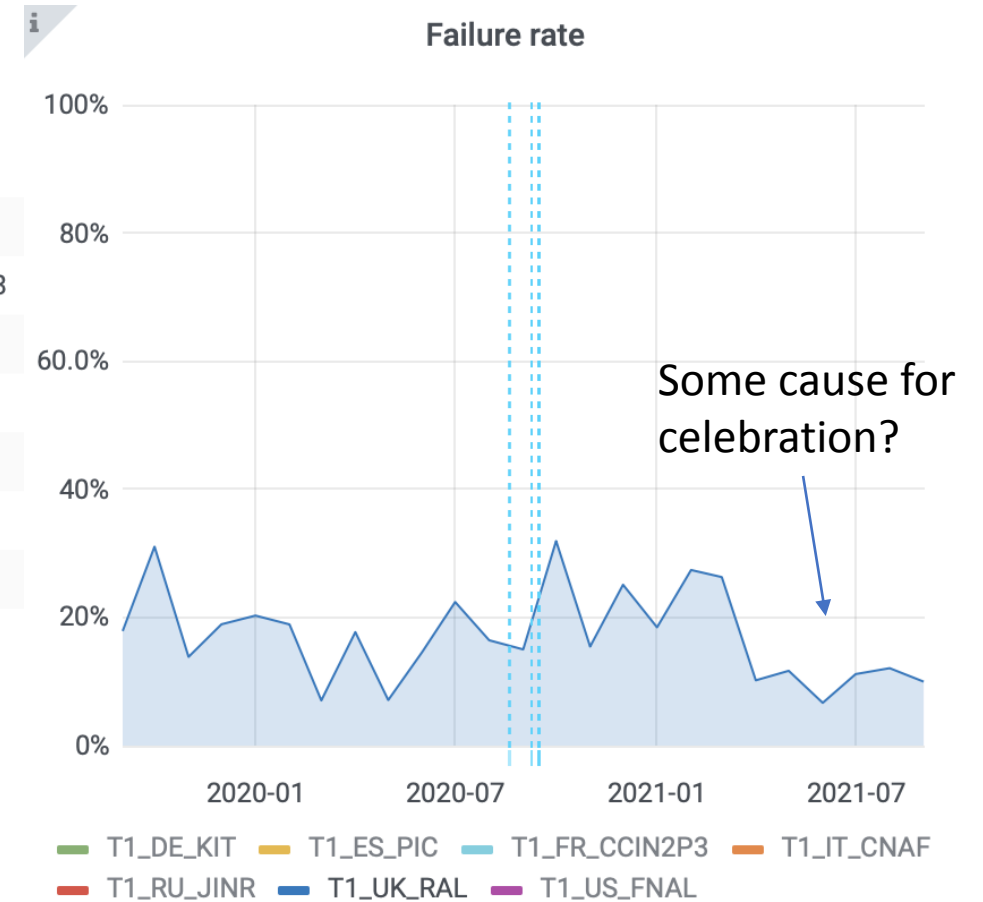
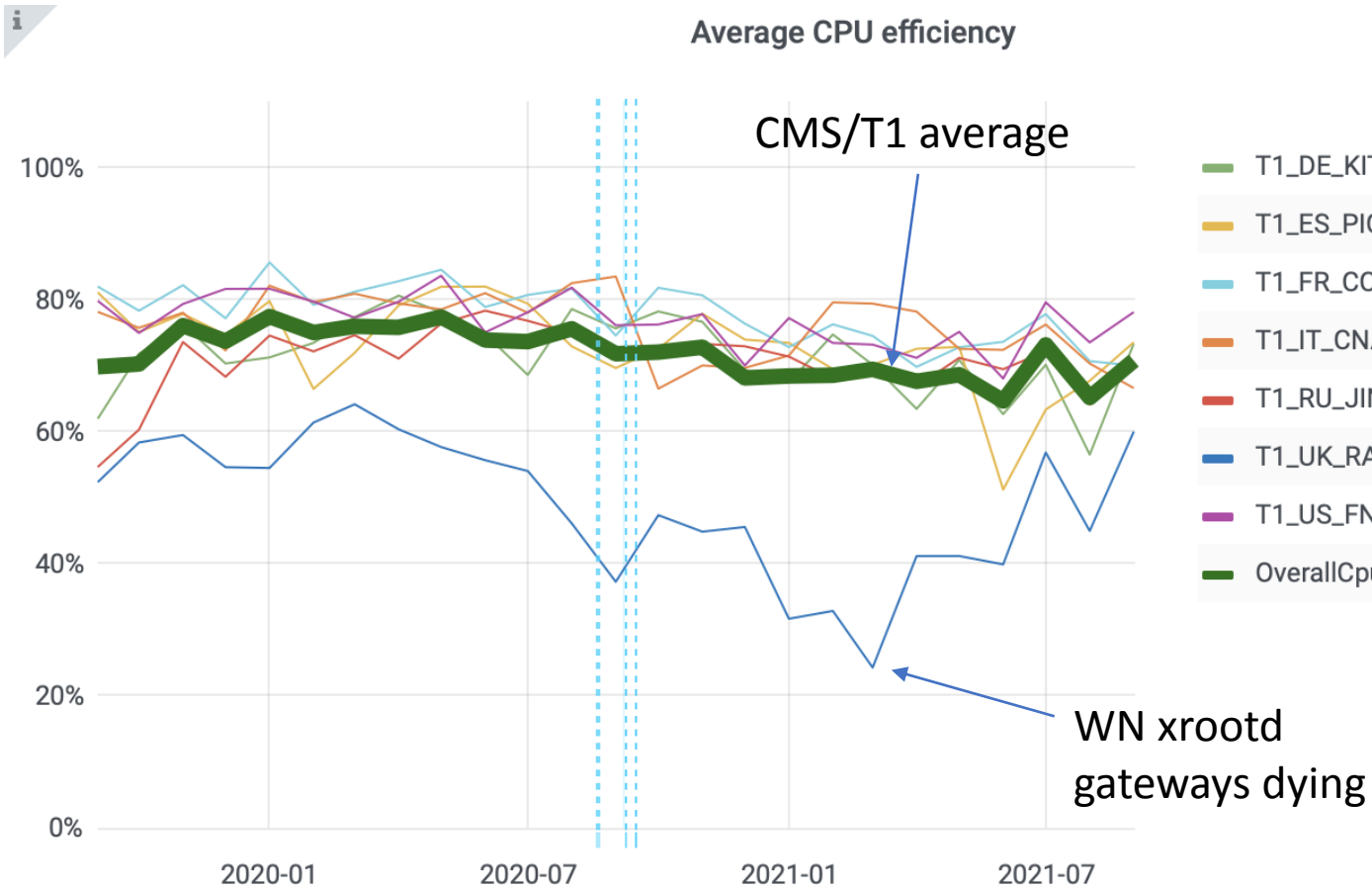


CMS job performance during a recent period at RAL

Katy Ellis, 02/09/21, GridPP46

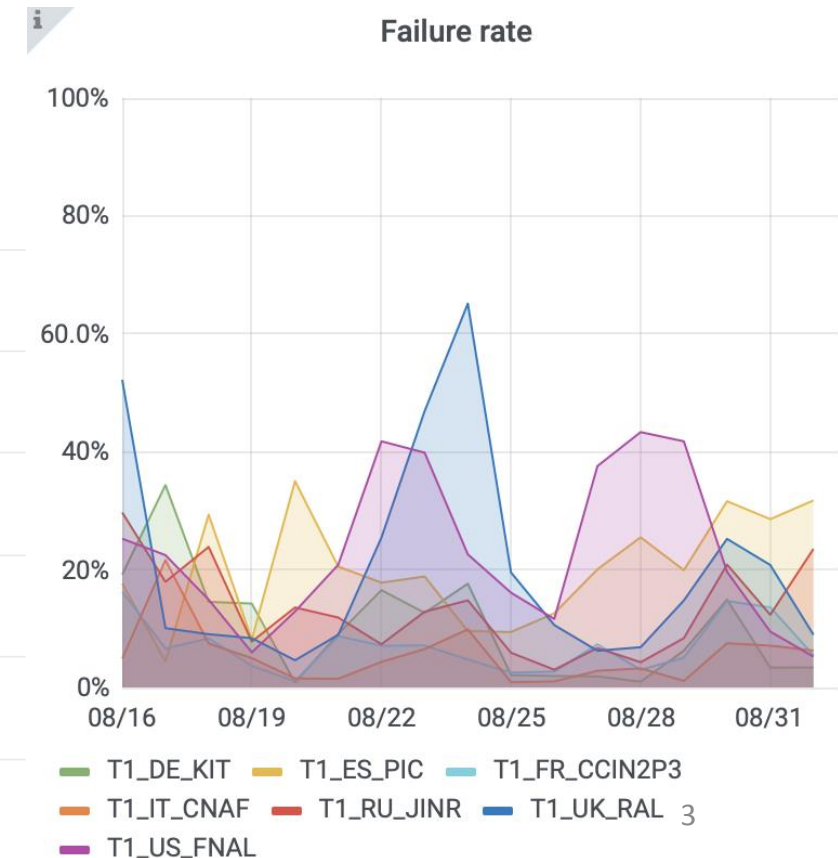
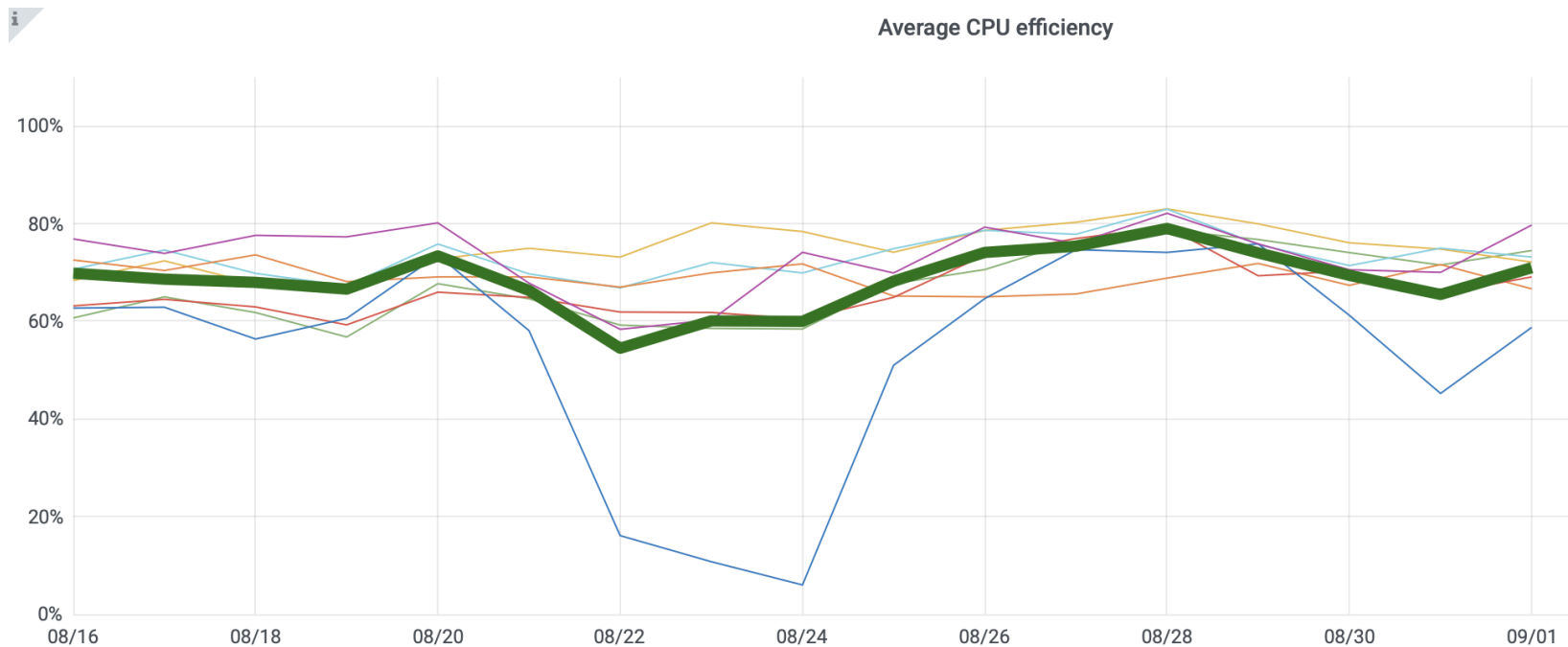
Summary of last 2 years



(CPU efficiency = CPU time / Core time)

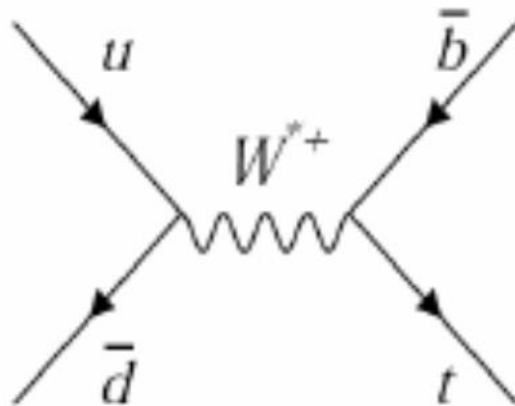
State of CMS jobs since recent changes

- Site core network change on weekend of 14/15 August
- Various WN upgrades in recent weeks/months
- Jobs reading from different locations?



First, a bit of context...

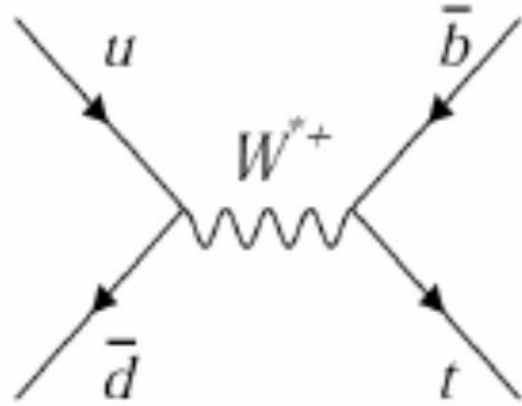
- An LHC 'event' is a snapshot of time in the detector
- The 'interesting' part is the 'hard scatter' – e.g., two particles have transferred a lot of energy between them
- However, there are a lot of other particles in the snapshot...



+

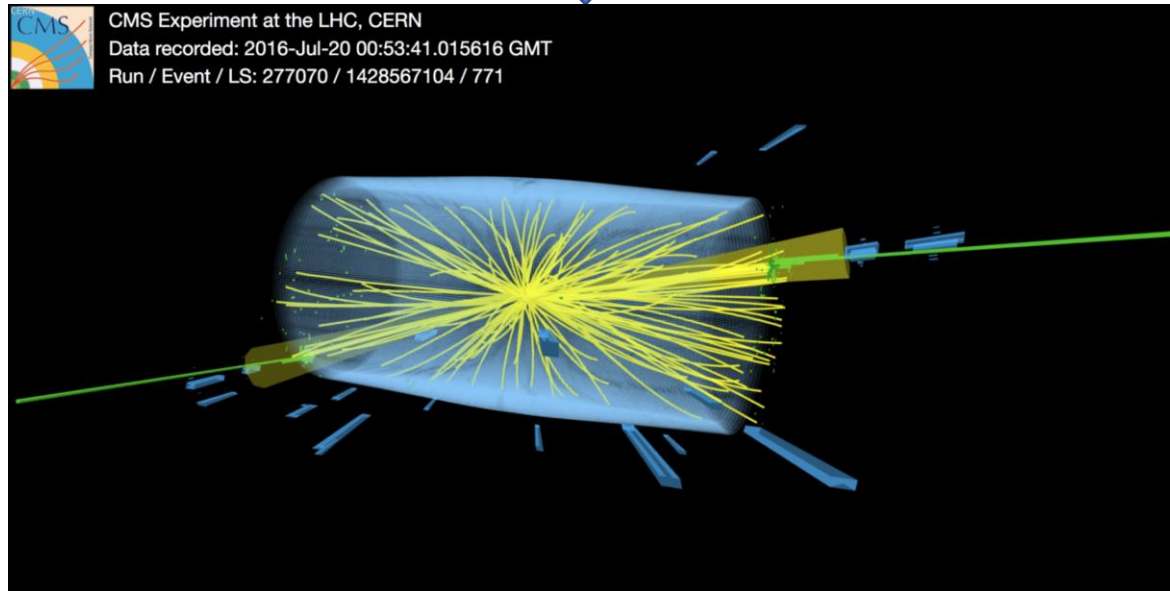
- 'Pile-up'
 - 'Leftover' parts of the proton
 - Radiation
 - Lower energy interactions from the same proton bunch

First, a bit of context...



+

- 'Pile-up'
 - 'Leftover' parts of the proton
 - Radiation
 - Lower energy interactions from the same proton bunch



First, a bit of context...

- When we simulate events, the **pile up or background events must also be simulated**.
- In CMS, a lot of the time we do this by generating pile up interactions separately and then overlay them on the main event
- Each main event might have 10s of pile up interactions
- The separate pile up data are stored in **huge datasets** of up to 700TB(?).
- These **datasets are located** typically **at CERN and FNAL** only
- Jobs at other sites access them via the **CMS AAA** service (remotely – aka ‘Offsite read’)

Monitoring Method

- Look at all jobs running at RAL T1 during some period
 - Some use 'secondary' inputs (pile up events); others do not.
 - List all the 'secondary' datasets and look up location(s)
 - Group the jobs by location of the secondary datasets (if any)
 - None, Onsite (RAL), Offsite (e.g. CERN), Offsite Outside Europe (e.g. USA)
 - Make plots of performance – failures and CPU efficiency
- N.B. my assumption is that data is accessed from the 'nearest' site...but I do not know this for sure. However, I am fairly confident that Onsite data comes from RAL T1 disk storage.

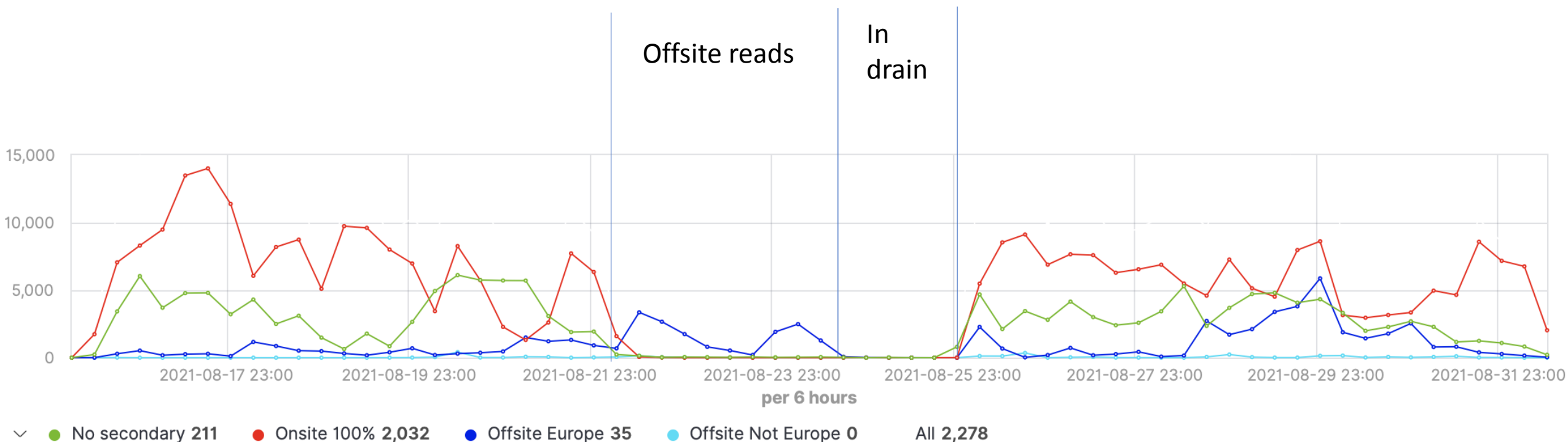
Locations of pile-up datasets in use

- /Neutrino_E-10_gun/RunIISpring15PrePremix-PUMoriond17_80X_mcRun2_asymptotic_2016_TracheIV_v2-v2/GEN-SIM-DIGI-RAW
 - /Neutrino_E-10_gun/RunIISummer20ULPrePremix-UL16_106X_mcRun2_asymptotic_v13-v1/PREMIX
 - /MinBias_TuneCP5_14TeV-pythia8/Run3Winter21GS-112X_mcRun3_2021_realistic_v15-v1/GEN-SIM
 - /Neutrino_E-10_gun/RunIISummer20ULPrePremix-UL17_106X_mc2017_realistic_v6-v3/PREMIX
 - /Neutrino_E-10_gun/RunIISummer20ULPrePremix-UL18_106X_upgrade2018_realistic_v11_L1v1-v2/PREMIX
 - /Neutrino_E-10_gun/RunIISummer17PrePremix-PUAutumn18_102X_upgrade2018_realistic_v15-v1/GEN-SIM-DIGI-RAW
 - /Neutrino_E-10_gun/RunIISummer17PrePremix-MCv2_correctPU_94X_mc2017_realistic_v9-v1/GEN-SIM-DIGI-RAW
 - /Neutrino_E-10_gun/RunIIFall17FSPrePremix-PUMoriond17_94X_mc2017_realistic_v15-v1/GEN-SIM-DIGI-RAW
 - /Neutrino_E-10_gun/RunIISummer16FSPremix-PUMoriond17_80X_mcRun2_asymptotic_2016_TracheIV_v4-v1/GEN-SIM-DIGI-RAW
- Offsite (70%CERN, 100%JINR)
 - Onsite
 - Onsite
 - Offsite (100%CERN)
 - Offsite (100%CERN)
 - Offsite (Not EU)
 - Offsite (Not EU)
 - Offsite (100%CERN)
 - Offsite (KIT, JINR, Purdue)

A quick note about the onsite data

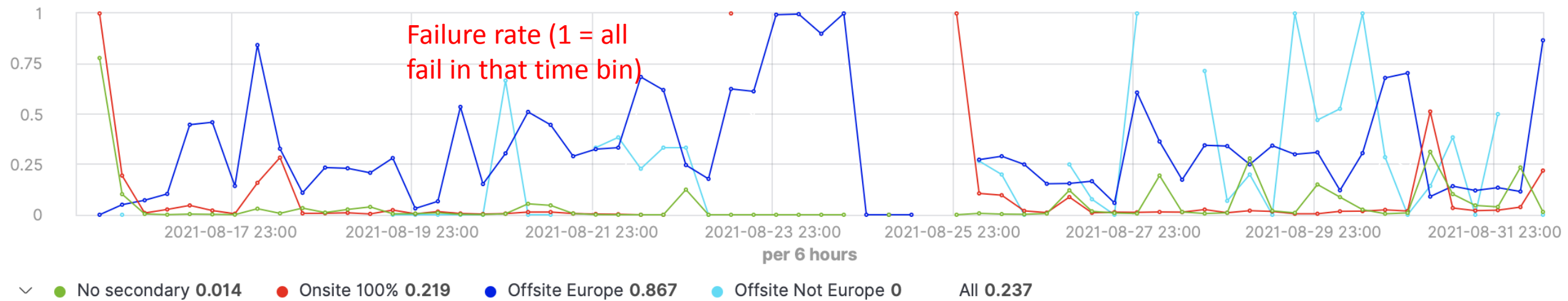
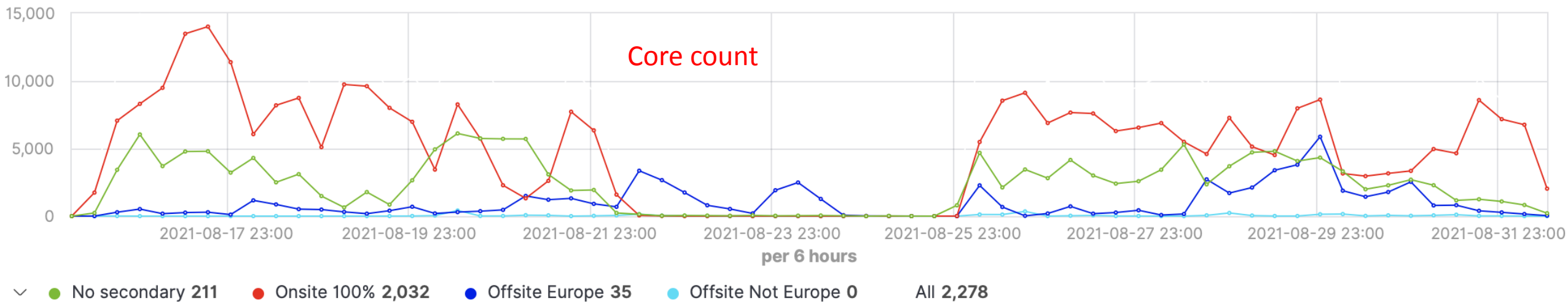
- The 'Neutrino_E-10...' onsite data was placed at RAL to test whether onsite reads were ok, and to see if we could improve the overall performance by having data more local
- The answer at the time (early 2021) was that it was no help – efficiency was still very bad
- I have a saved plot of this, but it's too confusing to show here, some information is missing, and the method of monitoring required inference about what the job was doing
- However, the dataset remains at RAL

CORE count

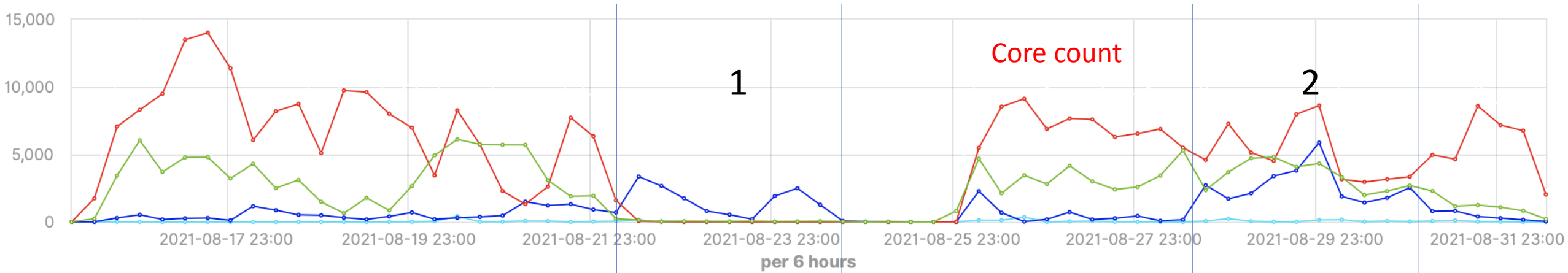


data.Status:Completed AND data.Site:T1_UK_RAL AND (NOT data.CMS_JobType:Analysis) AND (NOT data.CMS_JobType:Merge) AND (NOT data.CMS_JobType:LogCollect) AND (NOT data.CMS_JobType:Cleanup) 10

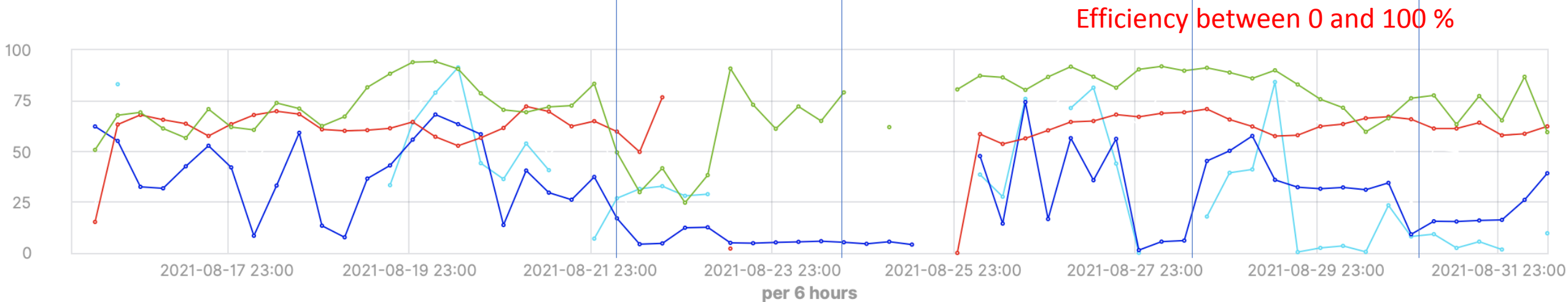
Job failures



Job efficiency (including failures)

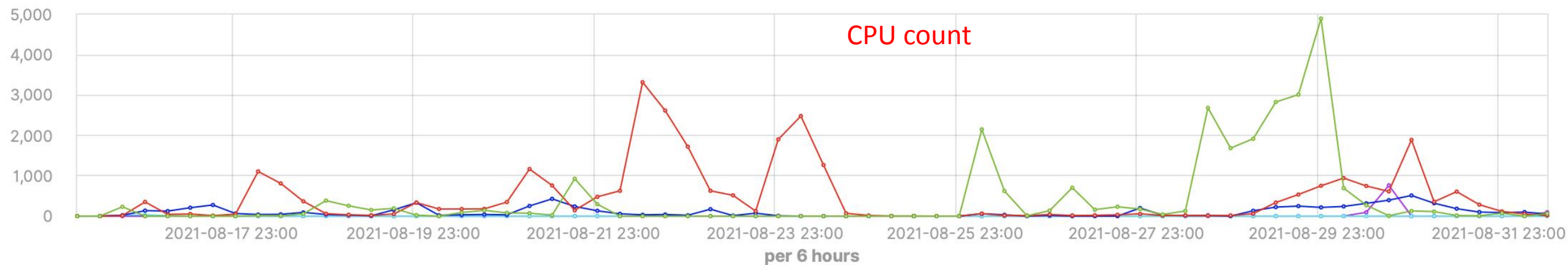


✓ ● No secondary 211 ● Onsite 100% 2,032 ● Offsite Europe 35 ● Offsite Not Europe 0 All 2,278

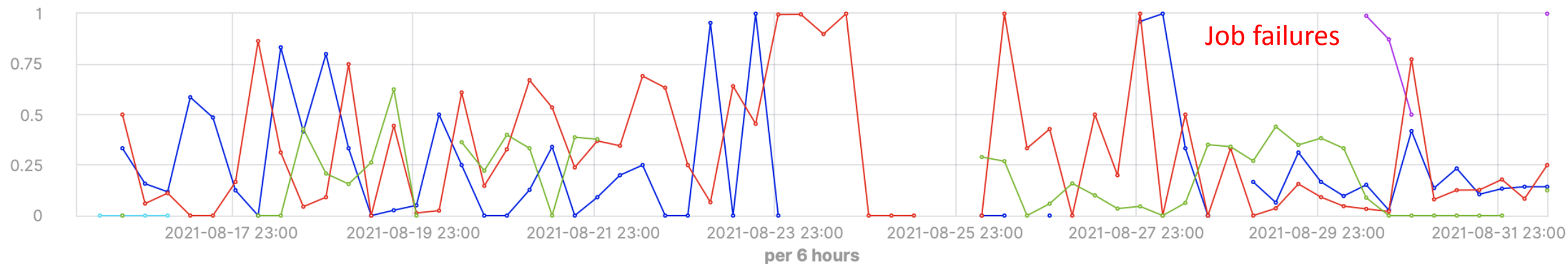


✓ ● No secondary 59.553 ● Onsite 100% 62.434 ● Offsite Europe 39.284 ● Offsite Not Europe 9.711 All 61.95

Further breakdown of Offsite reads (Europe)

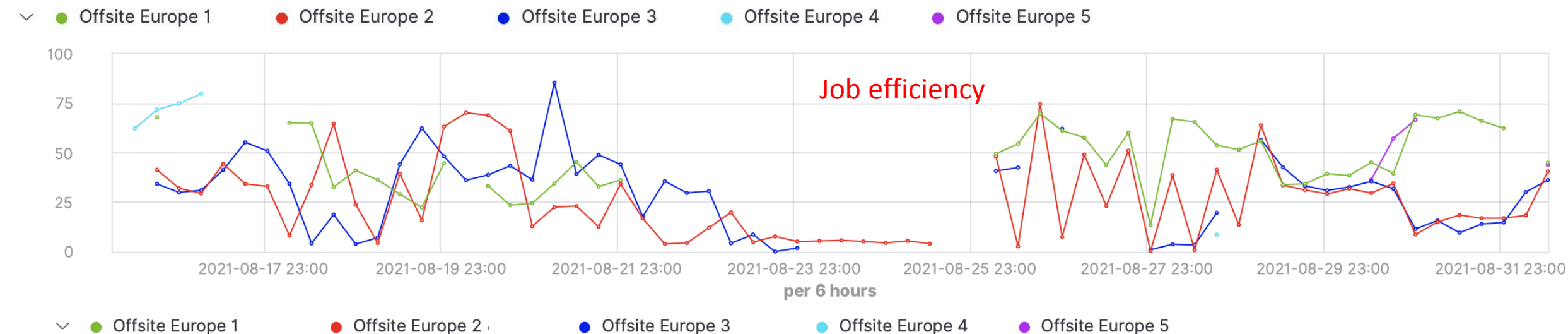
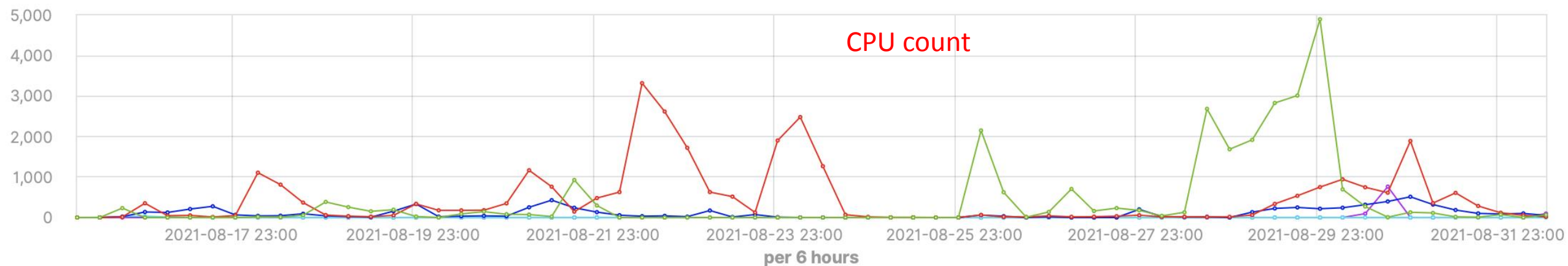


✓ Offsite Europe 1 Offsite Europe 2 Offsite Europe 3 Offsite Europe 4 Offsite Europe 5 ← Different pileup datasets




✓ Offsite Europe 1 Offsite Europe 2 Offsite Europe 3 Offsite Europe 4 Offsite Europe 5

Further breakdown of Offsite reads (Europe)



Deeper dive on 'Offsite Europe 2'

 **Data Aggregation System (DAS):** [Home](#) | [Services](#) | [Keys](#) | [Bug report](#) | [Status](#) | [CLI](#) | [FAQ](#) | [Help](#)

results format: , results/page, dbs instance ,

site dataset=/Neutrino_E-10_gun/RunII Summer20ULPrePremix-UL17_106X_mc2017_realistic_v6-v3/PREMIX

[Show DAS keys description](#)

Showing 1—3 records out of 3.

Hopefully not
coming from here!

Site name: T1_US_FNAL_Disk
Block completion: **100.00%** Block presence: **100.00%** File-replica presence: **100.00%** Site type: **DISK** StorageElement: T1_US_FNAL_Disk
[Datasets](#) Sources: [combined](#) [show](#)

Almost certainly not
coming from here!

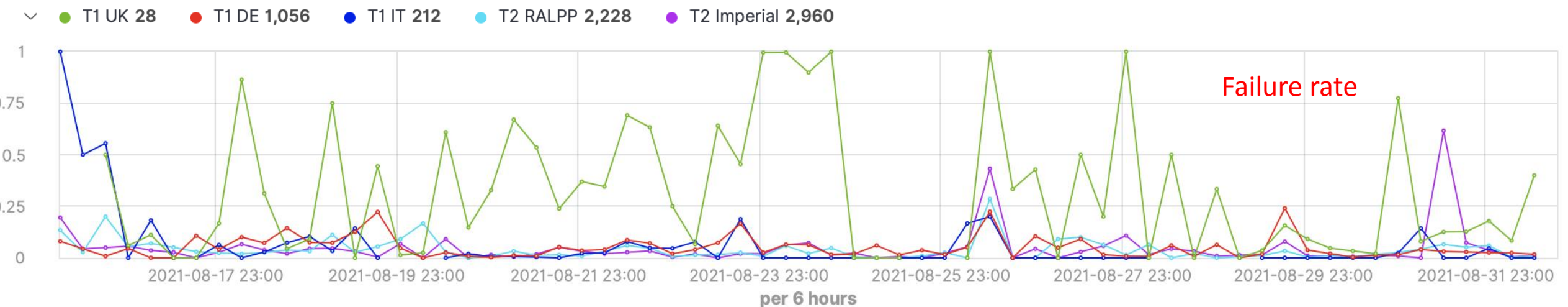
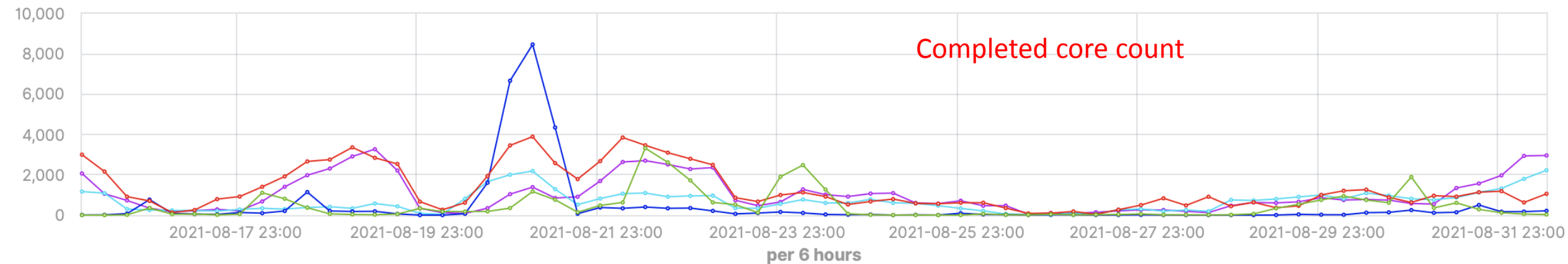
Site name: T1_US_FNAL_Tape
Block completion: **100.00%** Block presence: **100.00%** File-replica presence: **100.00%** Site type: **TAPE no user access** StorageElement: T1_US_FNAL_Tape
[Datasets](#) Sources: [combined](#) [show](#)

Probably coming
from here

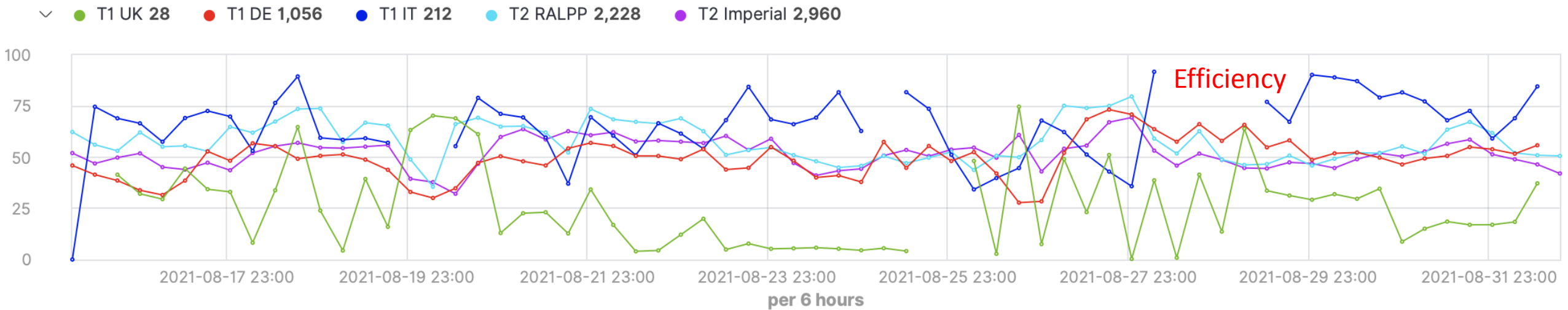
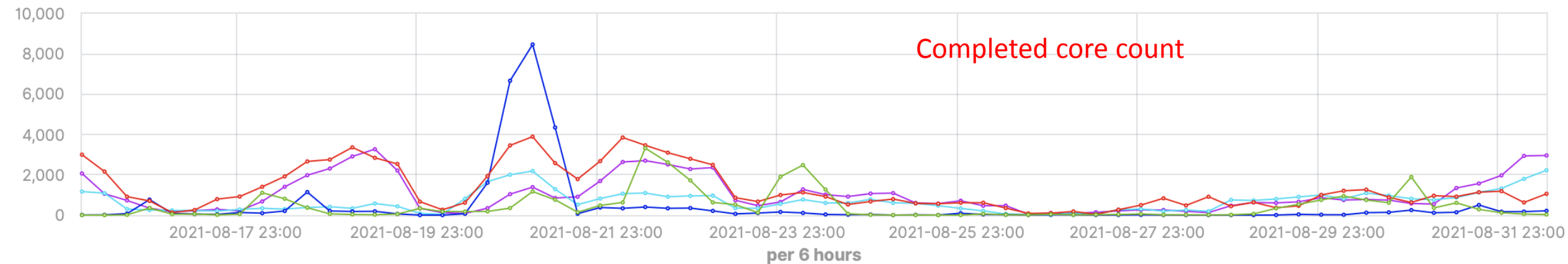
Site name: T2_CH_CERN
Block completion: **100.00%** Block presence: **100.00%** File-replica presence: **100.00%** Site type: **DISK** StorageElement: T2_CH_CERN
[Datasets](#) Sources: [combined](#) [show](#)

Showing 1—3 records out of 3.

Deeper dive on 'Offsite Europe 2'



Deeper dive on 'Offsite Europe 2'



Conclusions

- Splitting plots by input (secondary) datasets seems a sound method
- From the available information, it shows that no secondary/Onsite read jobs are doing quite well
- Offsite read jobs have by far worst performance at RAL
- Compared with other similar sites using the same offsite inputs, the failure rate is far higher, and the CPU efficiency is far lower
- Looking forward to the T1 network ~~improvements~~ complete redesign coming in the next month or two to help improve remote data access

Backup

Input data

