CMS

(a checkpoint)

Adriano Di Florio (CC-IN2P3)

LCG France Meeting - 26 June 2025 - CC-IN2P3

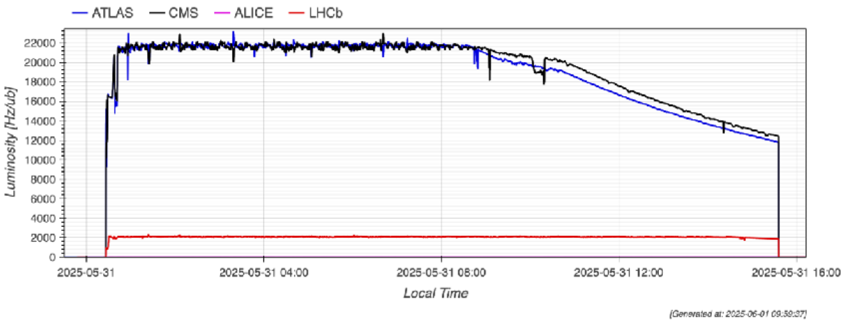
Disclaimer

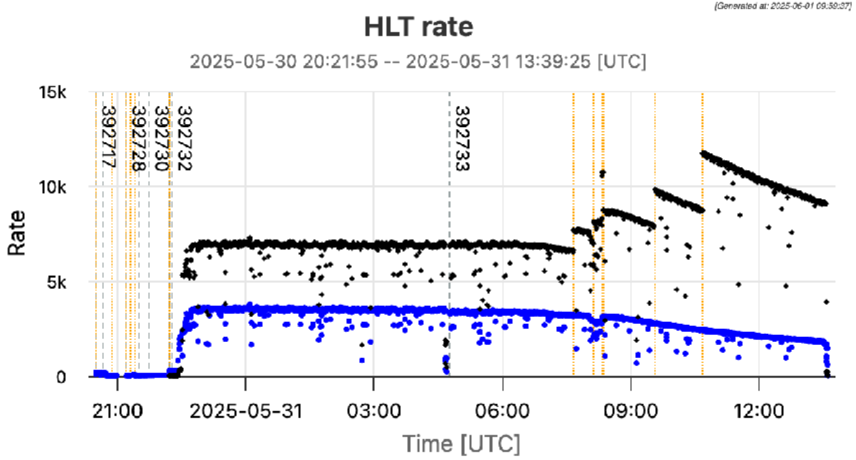
Ils sont me premierées Journées LCG France
en tant que CMS contact. Donc je passerais en
l'anglais.

But of course for any questions or discussion we can go back to French.

2025 data taking just started

May 31: CMS is entering the “more than 1 fb^{-1} of data” per fill «era»





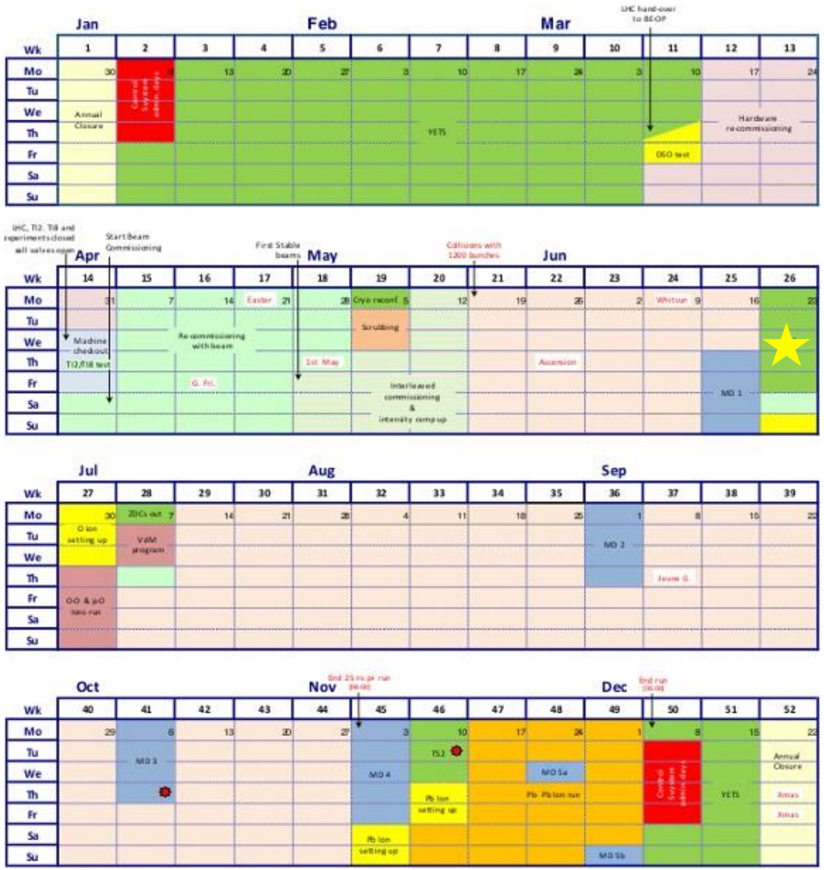
First fill with more than 1 fb^{-1} of data recorded

- Fill: 10676
- Date: March 31, 2025
- Duration (Stable Beams): 15.5 hours
- Delivered: 1.06 fb^{-1}
- Recorded: 1.01 fb^{-1}
- Efficiency: 95%
- Levelling at PU63
- L1 rate: 110/105 kHz
- Deadtime: 3-4%
- Levelling: 3.5 kHz (Physics stream) → 7 kHz total

2025 data taking just started

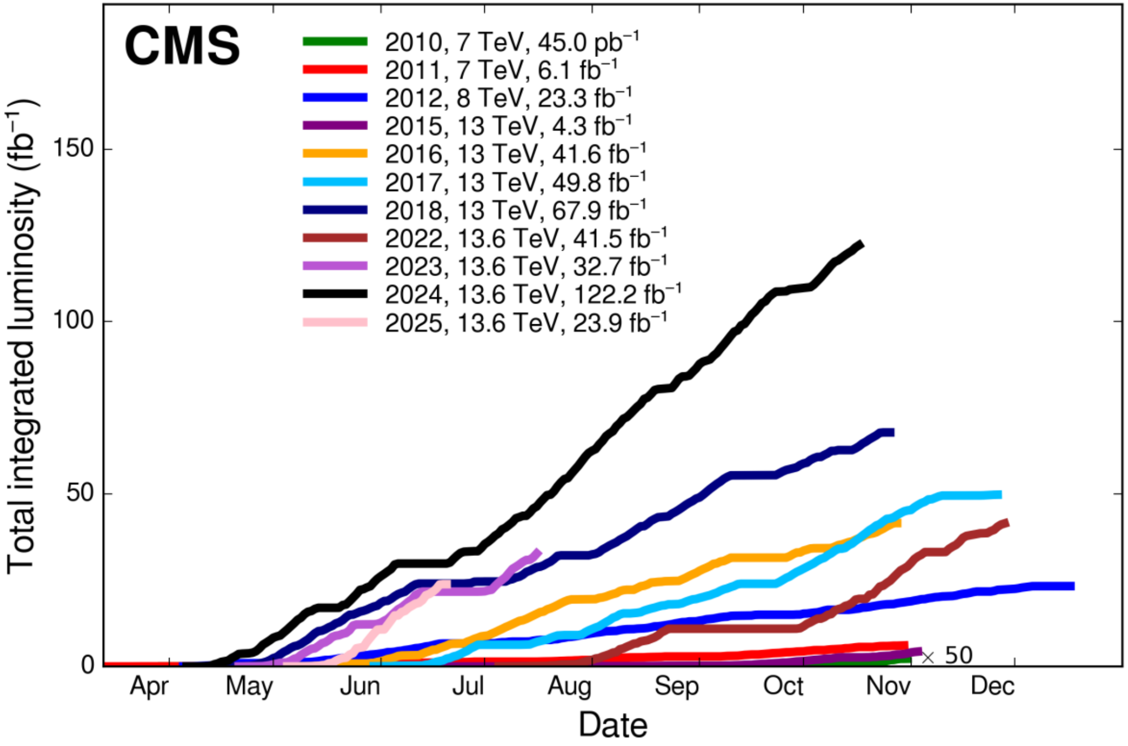
LHC Commissioning activities are on schedule

- 8th Apr: start of LHC beam commissioning
  - 14th April: splashes
  - 22nd April: 900 GeV SB collisions
  - 2nd May: PPS BBA
- 5th May: first stable beams at 13.6 TeV
- 19th May: 1200 bunches in LHC → physics production
- 19th - 22nd Jun: MD1 (Machine Development)
- 23th - 27th Jun: TS1 (Technical Stop)
- 29th Jun - 6th Jul: pO + OO (+NeNe) run
- 8th - 9th Jul: VdM pp
- 1st - 4th Sep: MD2
- 6th - 9th Oct: MD3
- 3rd - 7th Nov: MD4
- 8th Nov: PbPb commissioning
- 10th - 12th Nov: TS2
- 15th Nov - 6th Dec: PbPb run



2025 data taking just started

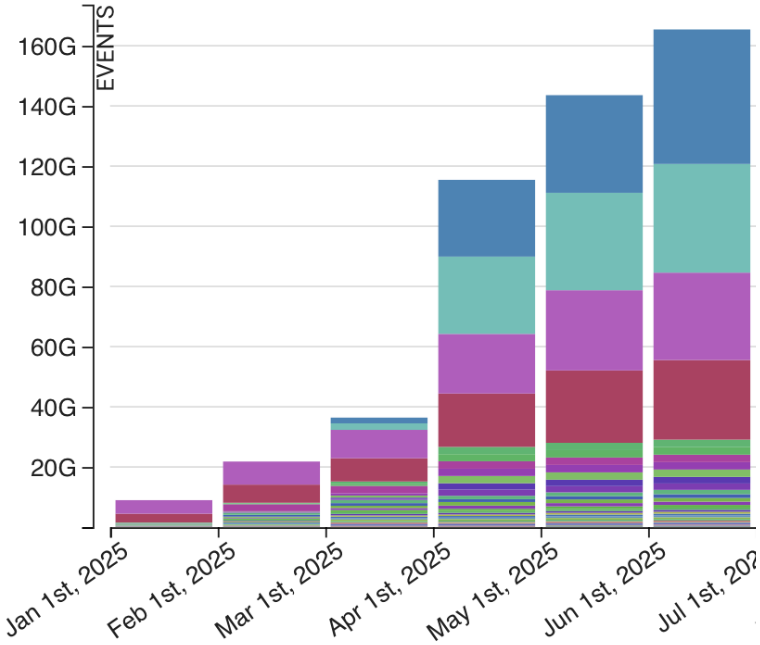
- ~ reached the targe of 25 fb^{-1} before the first MD



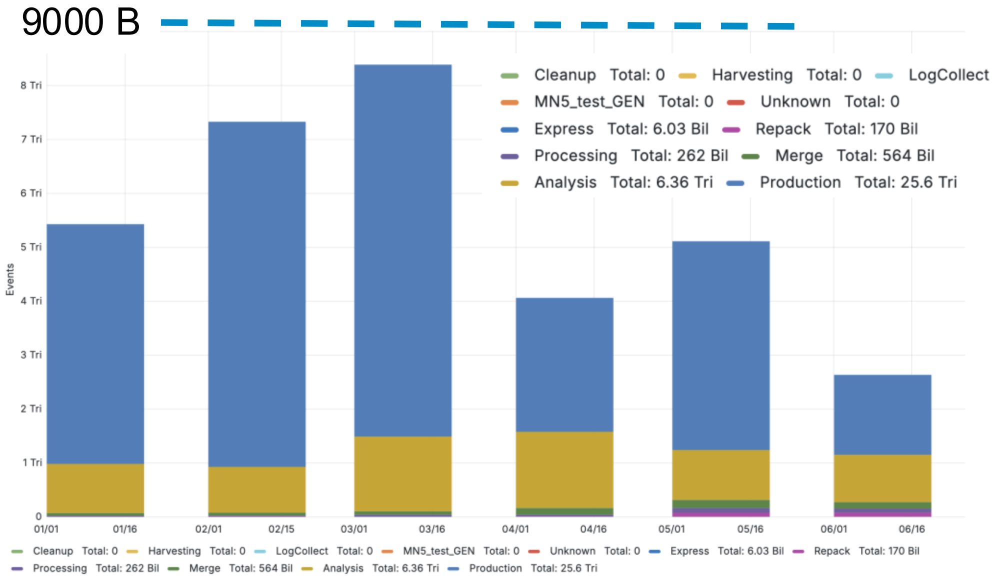
Target integrated luminosity : 120 - 150 fb^{-1} (at least as 2024)

2025 MC production

- Whole CMS MC production ~160B events written since the beginning of the year (~6B per week)



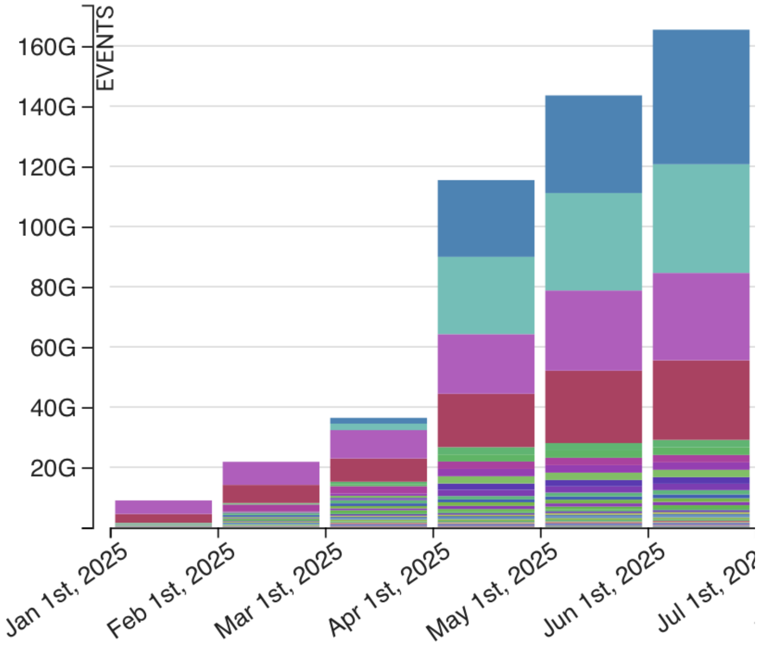
(injected events x generation efficiency)



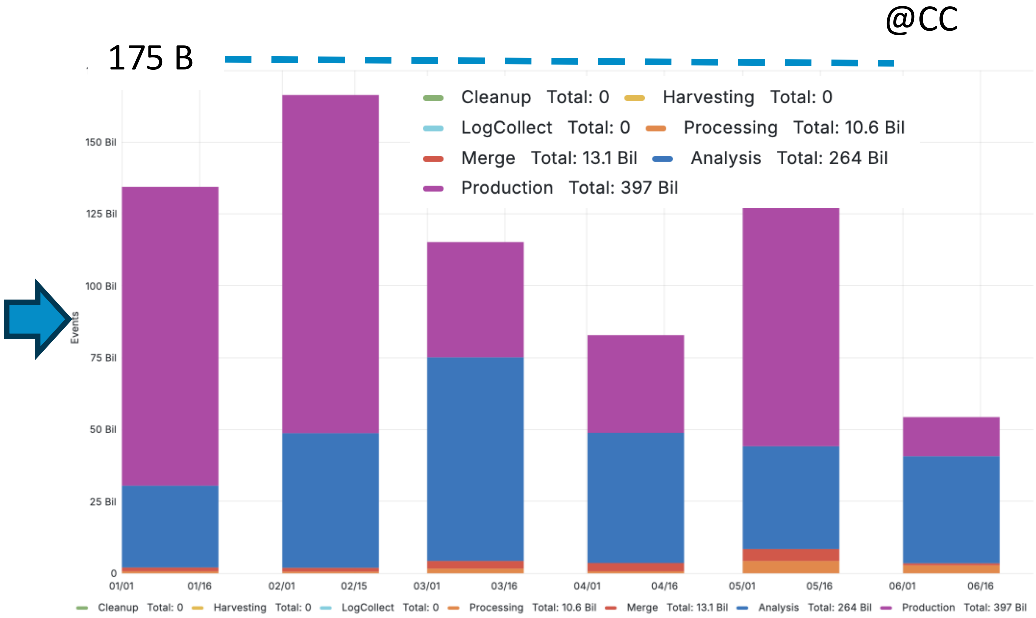
(injected events)

2025 MC production

- Whole CMS MC production ~160B events written since the beginning of the year (~6B per week)

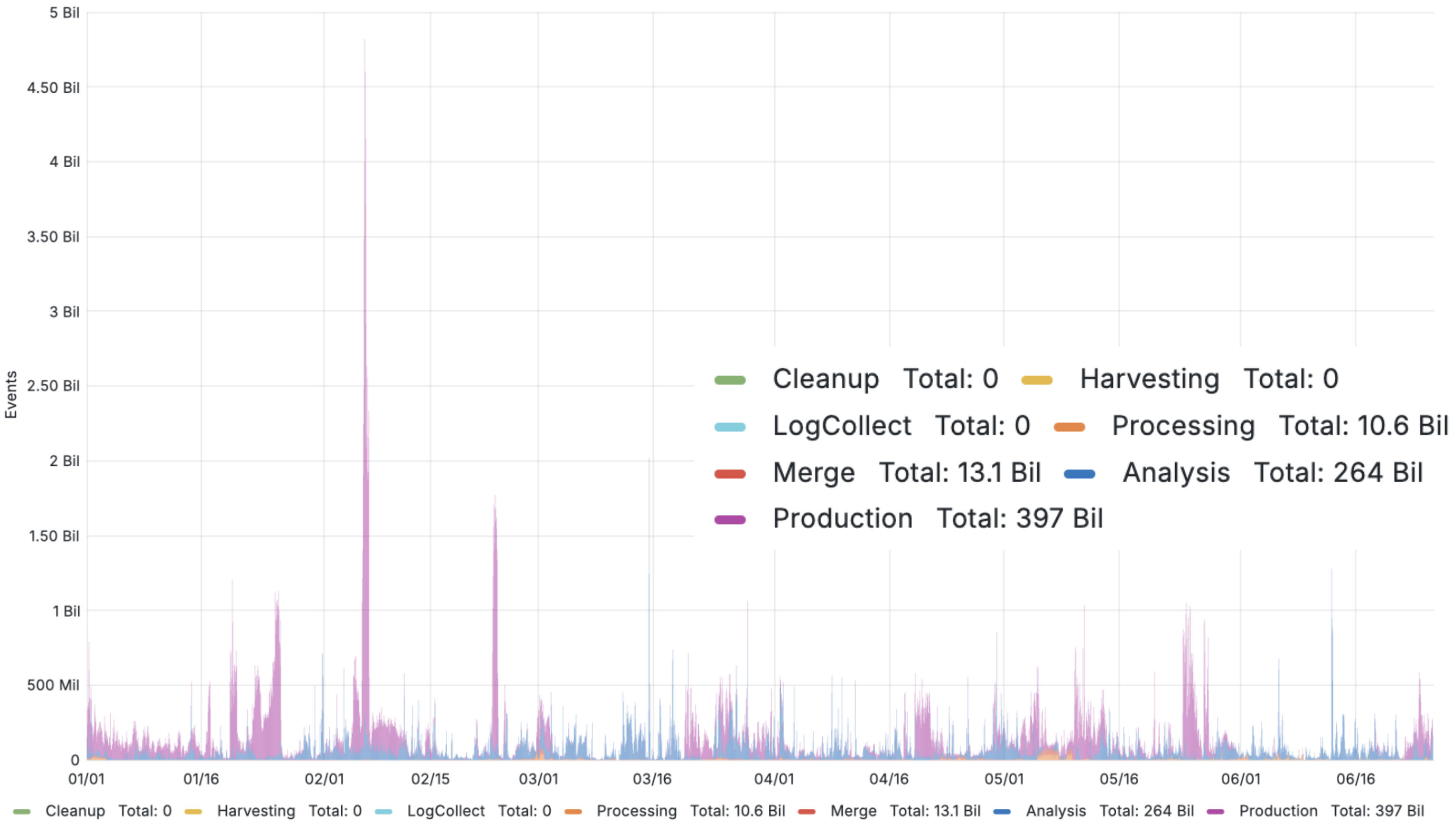


(injected events x generation efficiency)

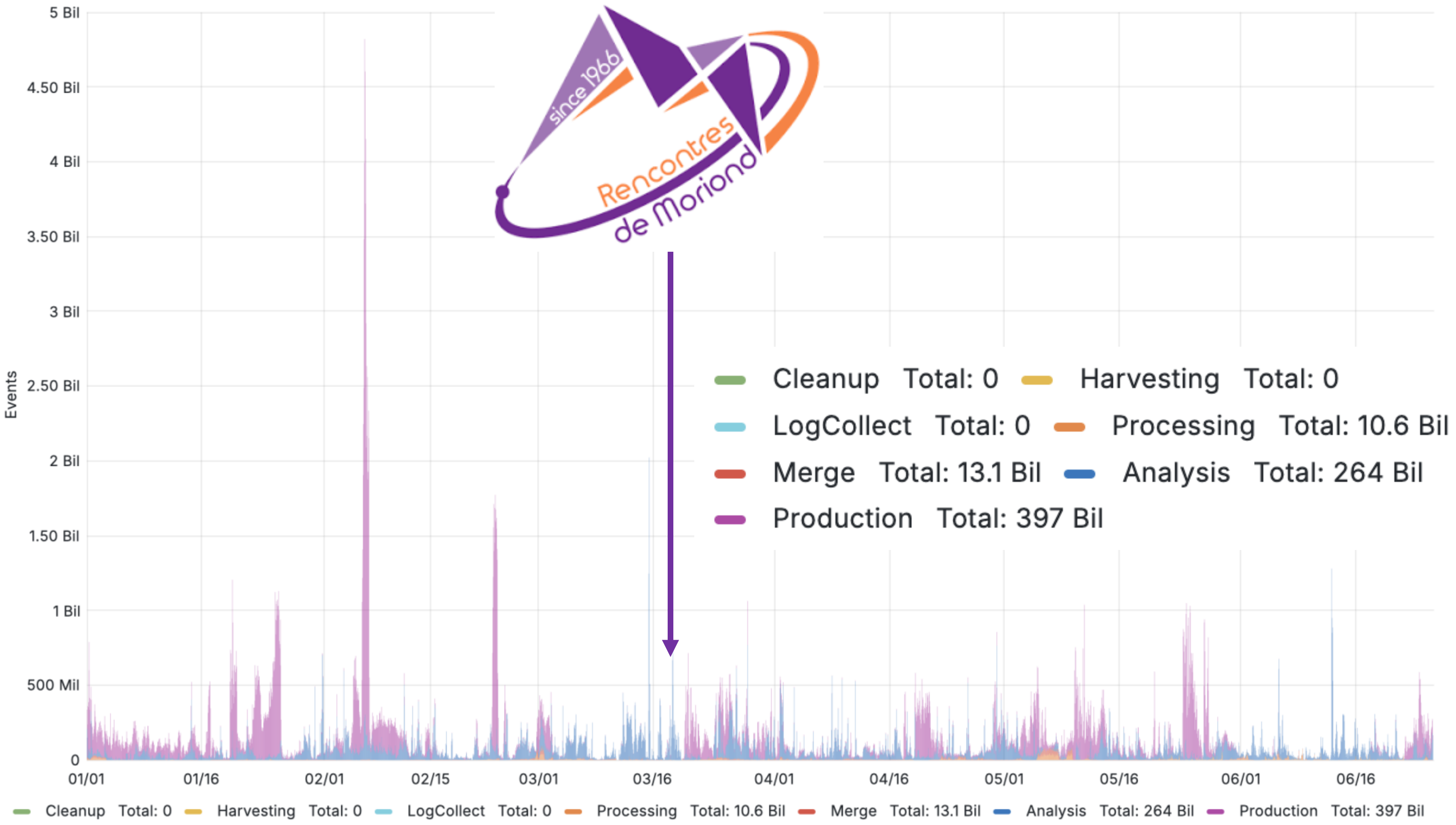


(injected events) ~700 B

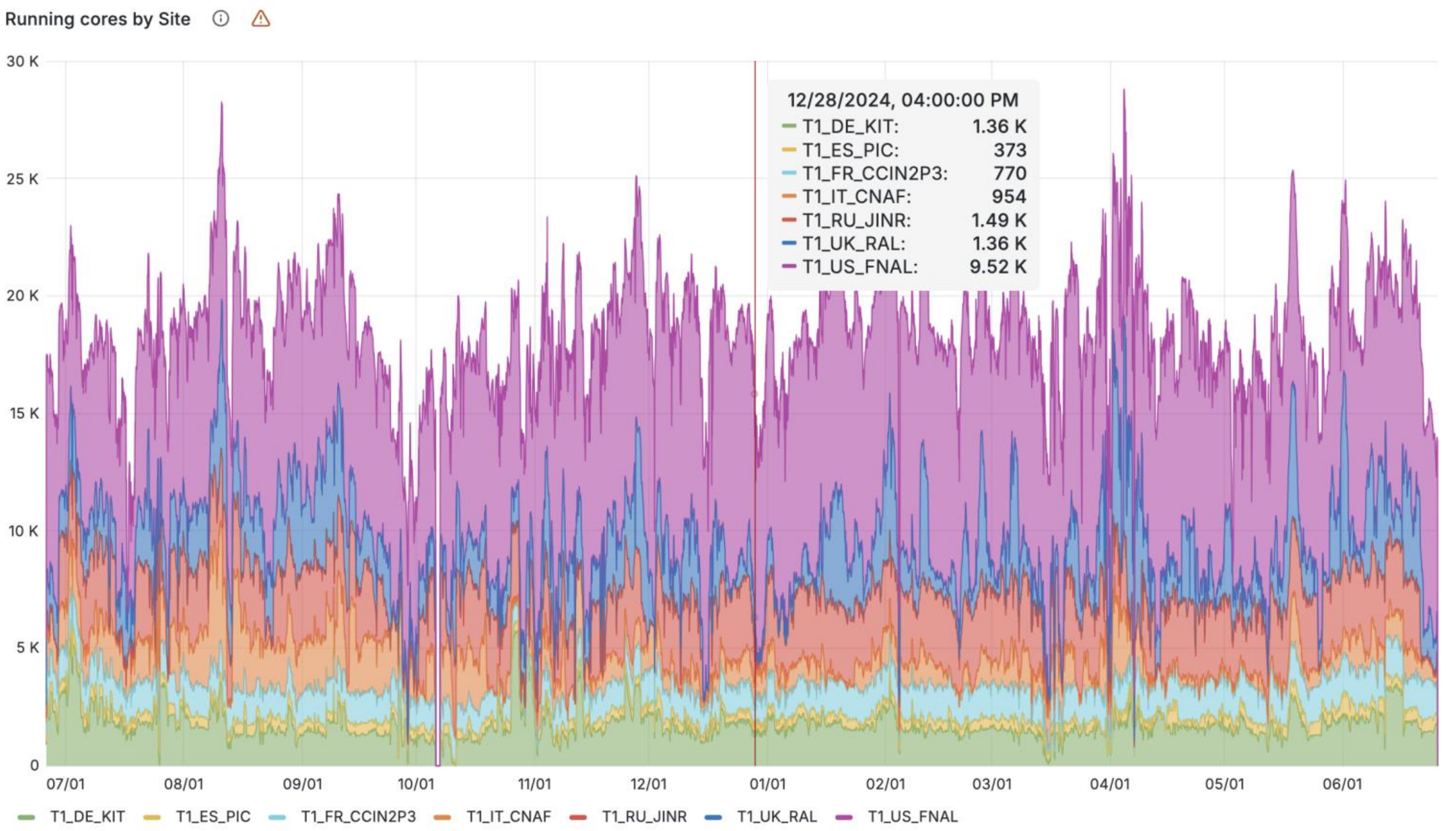
2025 MC Production – Peaks @ CC



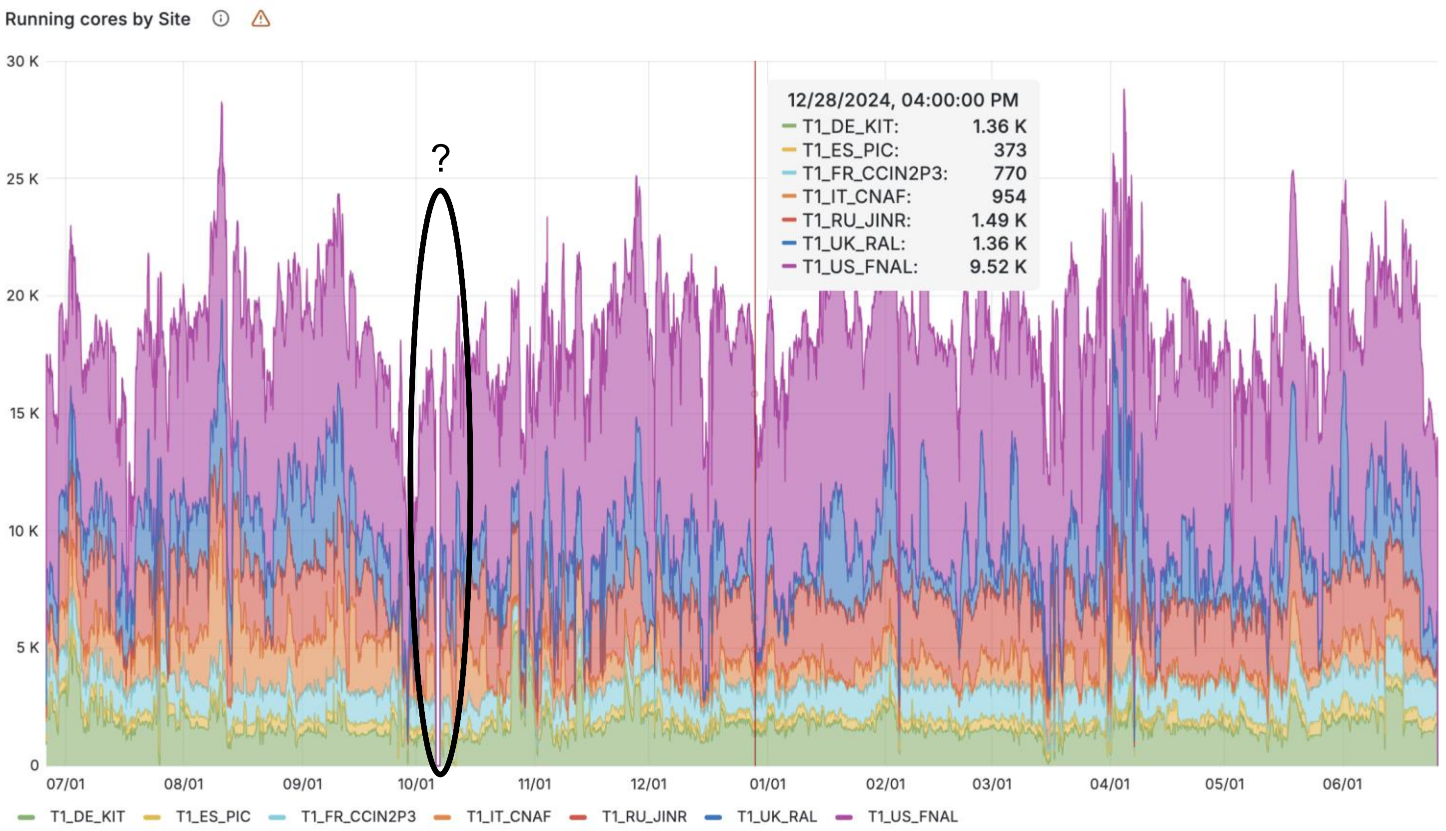
2025 MC Production – Peaks @ CC



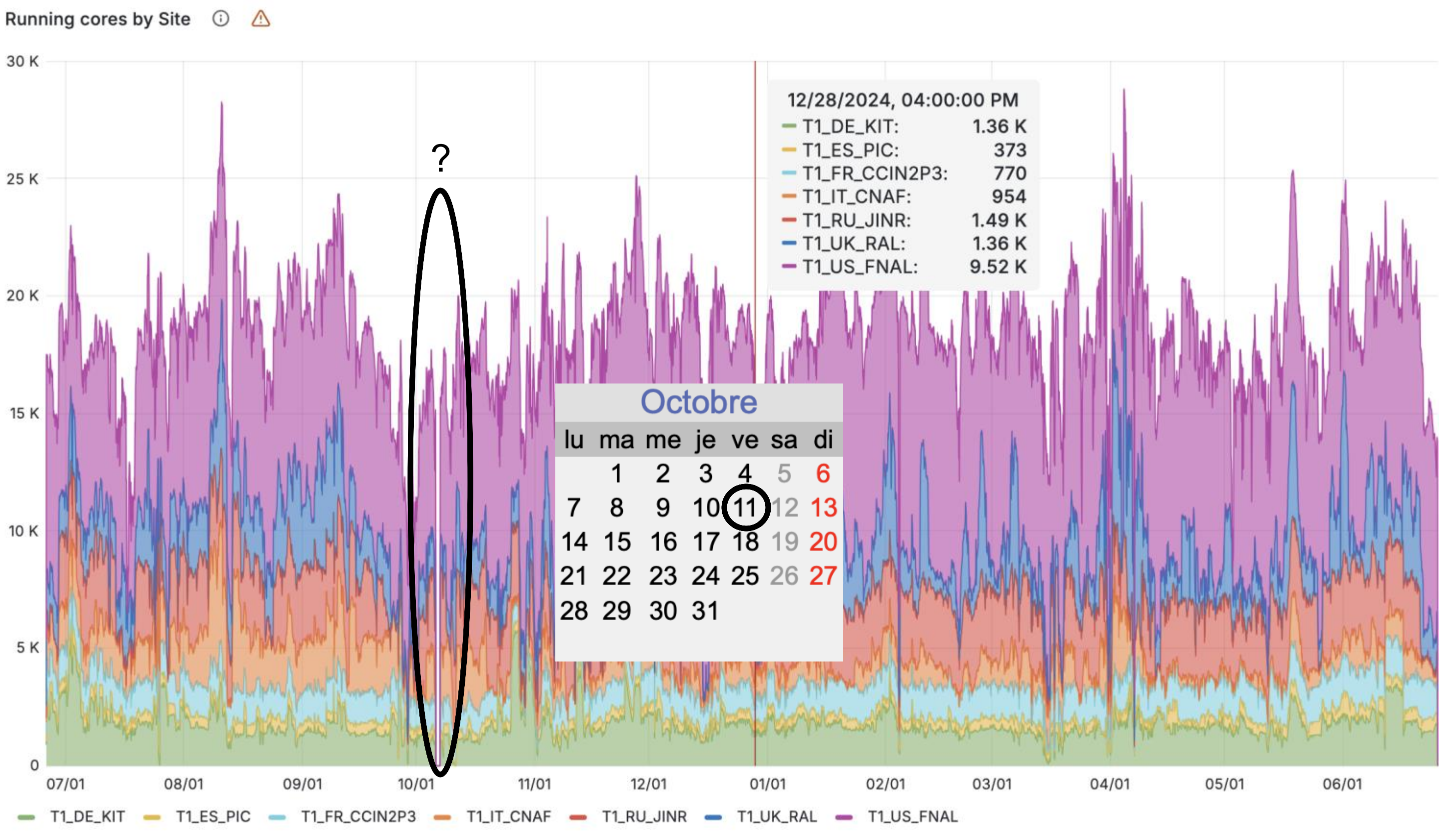
Running Cores @ T1s



Running Cores @ T1s



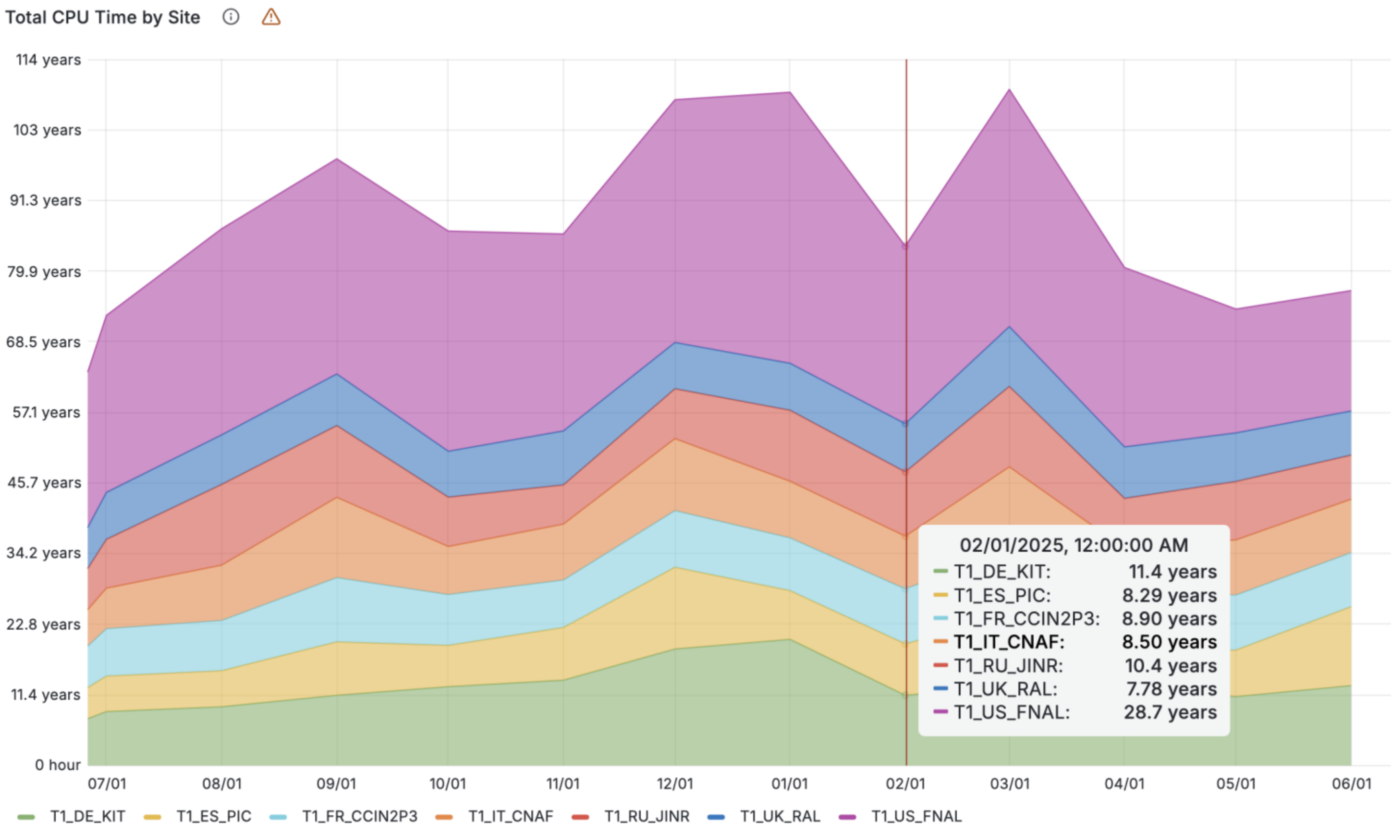
Running Cores @ T1s



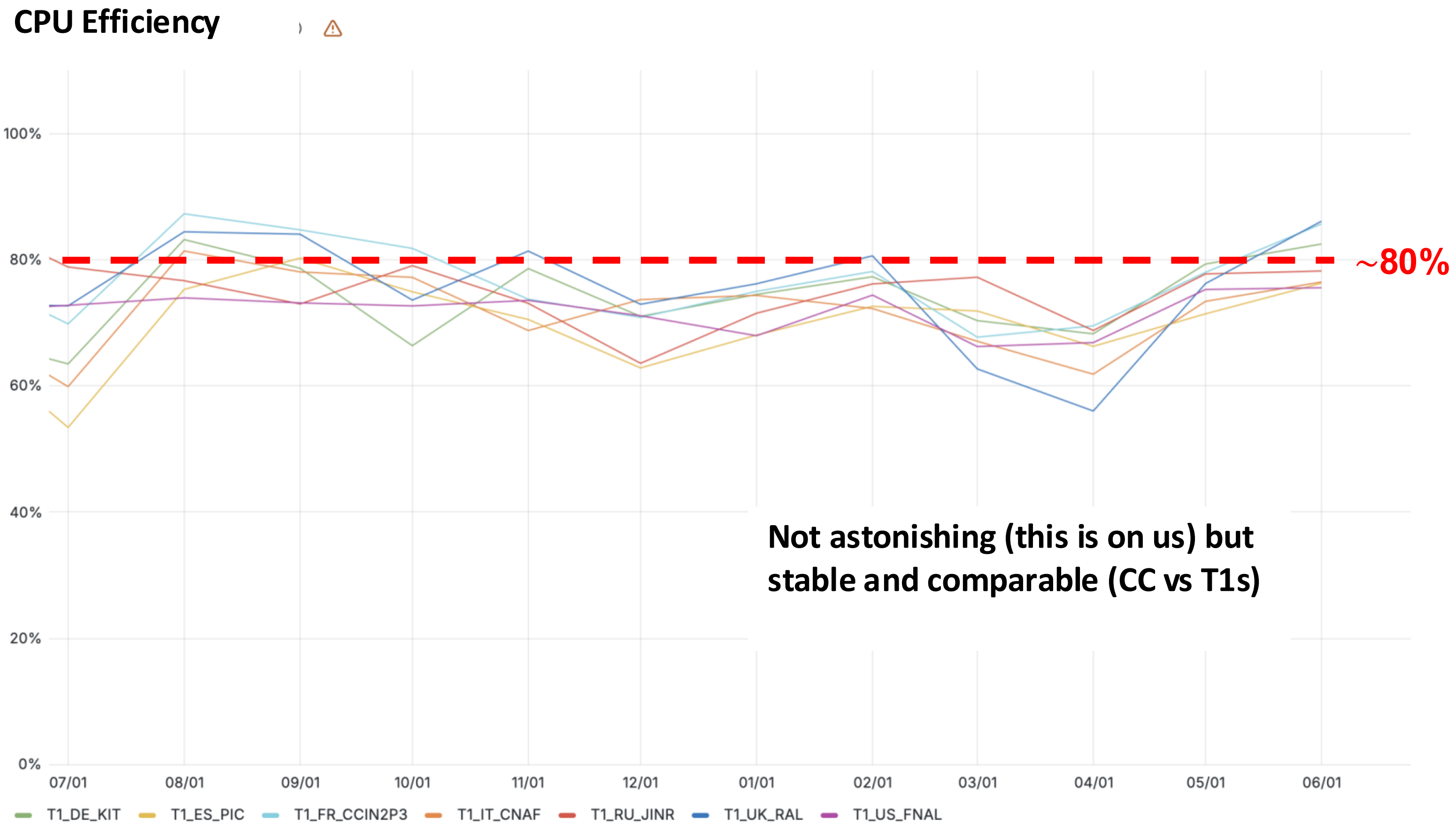
This is a Friday afternoon (buggy) patch for the HTCondor matchmaking mechanism.

CPU Time @ T1s

- T1_FR_CCIN2P3 stable in the 10-15% range

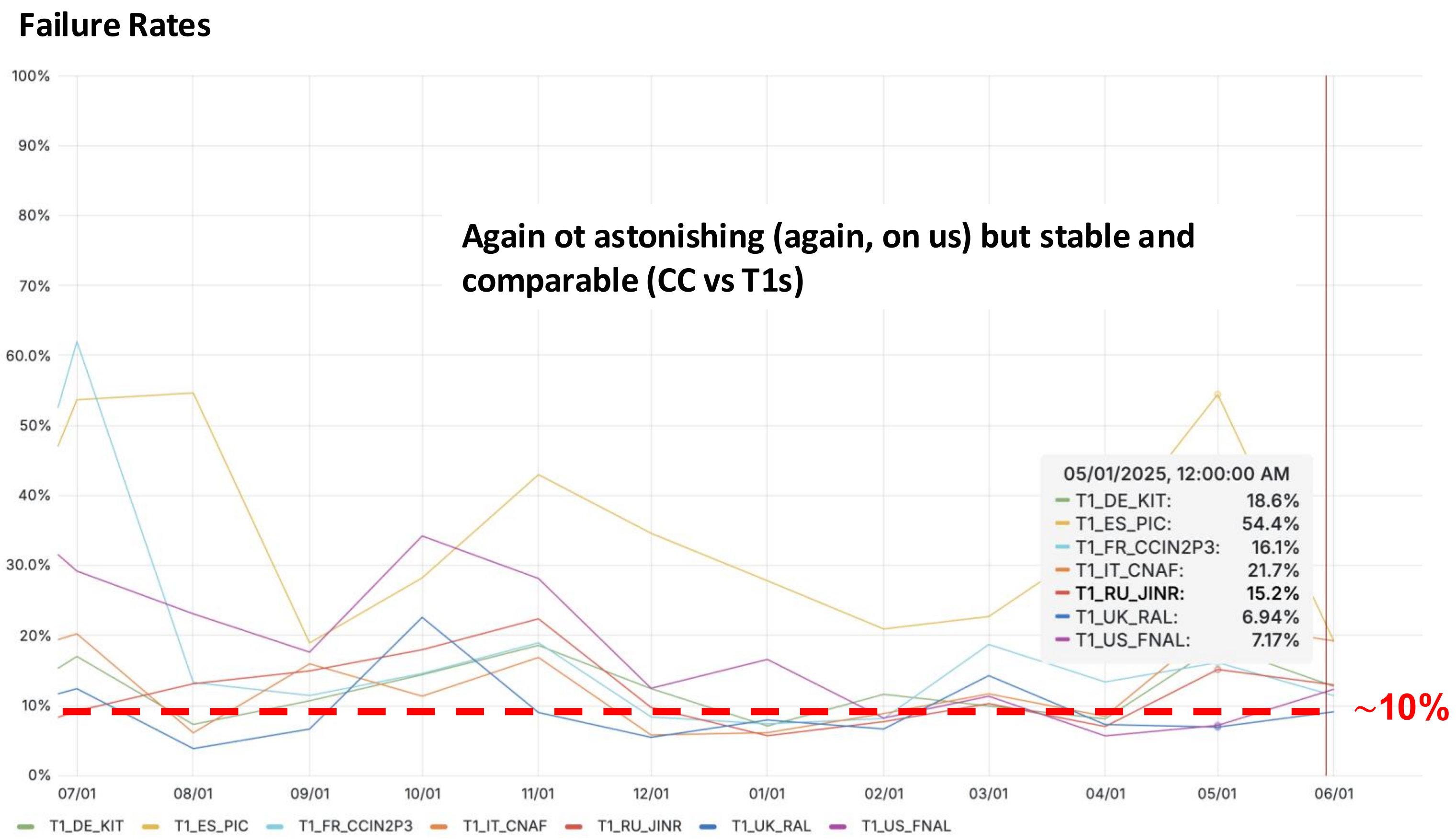


CPU Efficiency (@T1)

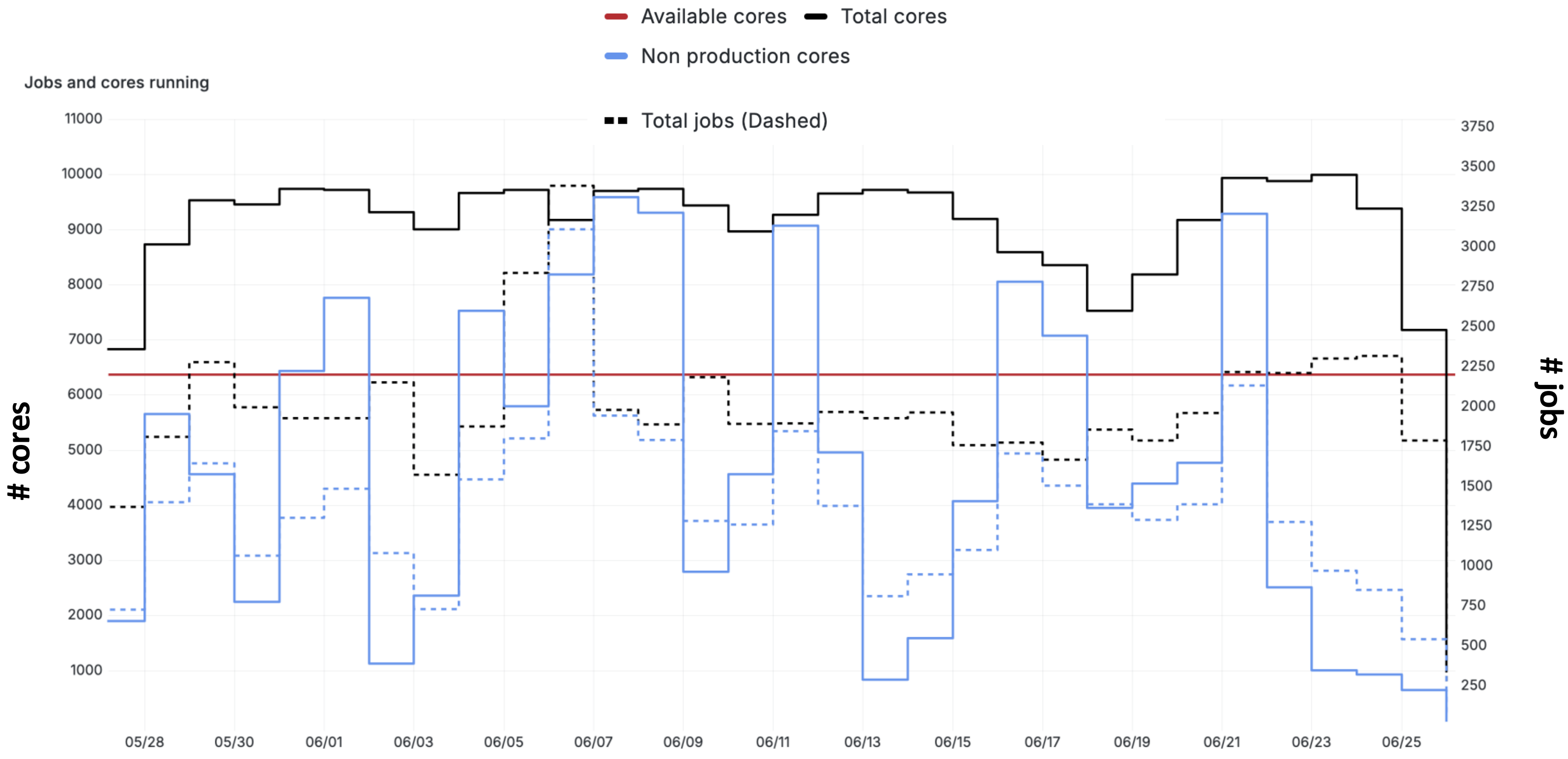


Not astonishing (this is on us) but stable and comparable (CC vs T1s)

Failure rates (@T1)



One month of jobs running (@CC)



2025-2026 Planning

- After LHC scenario was updated for Spring 2025 C-RSG report
  - The collaboration asked to increase the trigger rates for both prompt (slightly), and parking (~1kHz) to allow more acceptance in di-Higgs, flavour physics, and new physics searches
  - We finalized 2026 resource request with new trigger rate
  - For 2025, we will manage the resource internally
    - Tier-0 & Tier-1: Tape clean up, mostly on pre-UL Run 2,
    - Tier-1: Reducing amount of AODSIM; now saved only on demand, i.e. store only MiniAOD, and NanoAOD outputs by default

Fall 2024 C-RSG

Parameter	2025	2026*
LHC		
LHC Energy pp [TeV]	13.6	
Average (Peak) pileup	62 (65)	
Integrated luminosity / year [fb^{-1}]	120	
Livetime pp / year [s/10^6]	6.3	
Livetime HI / year [s/10^6]	1.4	
Heavy Ion run type	Pb-Pb,O-O,Pb-O	
CMS-Specific		
Prompt HLT Rate [kHz]	2.6	
Parked HLT Rate [kHz]	4.9	
HLT Scouting Rate [kHz]	30	
L1 Scouting Rate [Orbits/sec]	1100	
Run 3 MC events / year in billions	57	
Phase-2 MC events / year in billions	0.5	

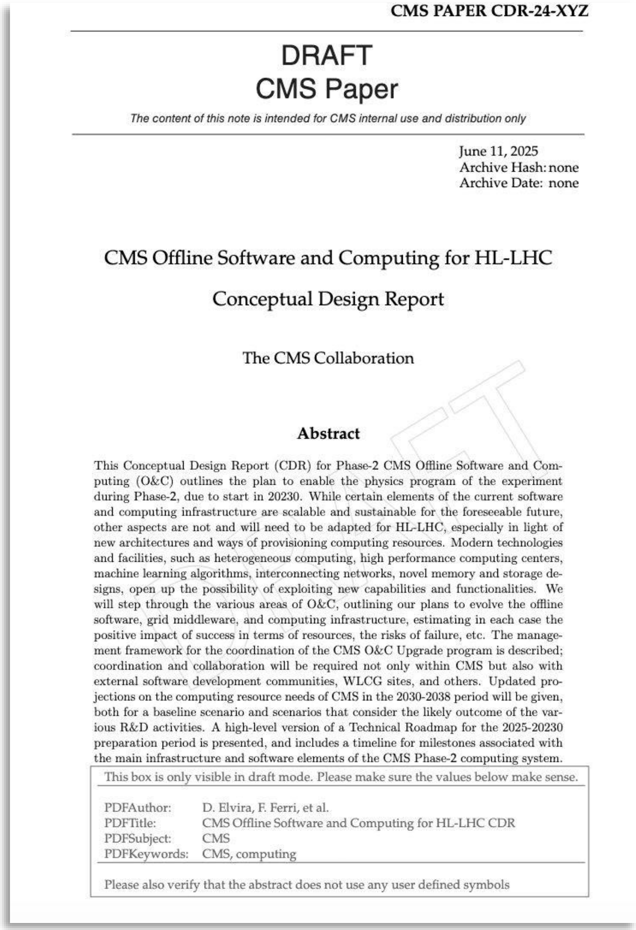
Spring 2025 C-RSG

Parameter	2025	2026
LHC		
LHC Energy pp [TeV]	13.6	
Average (Peak) pileup	65 (67)	
Integrated luminosity / year [fb^{-1}]	150	80
Livetime pp / year [s/10^6]	6.5	3.5
Livetime HI / year [s/10^6]	1.2	1
Heavy Ion run type	Pb-Pb,O-O,Pb-O	
CMS-Specific		
Prompt HLT Rate [kHz]	2.8	2.8
Parked HLT Rate [kHz]	6.0	6.0
HLT Scouting Rate [kHz]	30	
L1 Scouting Rate [Orbits/sec]	1100	
Run 3 MC events / year in billions	69	41
Phase-2 MC events / year in billions	0.5	

- Plan to use both Lustre, and SSD flows
- For resources;
  - Tier-0:
    - +122 PB of Tape
    - +16 PB of disk
    - Request addition of 25 PB of tape and 4 PB of disk to accommodate higher int. luminosity, and pile-up
  - Tier-1:
    - +58 PB of Tape (of +65 PB request),
    - Request additional of 9 PB of tape for higher statistics MC samples

Conceptual Design Report

- *The* document describing CMS plans for Phase2 for offline operations.
- Demonstrate at the conceptual level that the proposed Computing Model (CM) fulfills the requirements from the HL-LHC physics program in the context of the collider and CMS detector scenarios
- Provides new estimations updating the old 2022 ones.
- An LHCC review based on CDR will be held sometime late in 2025 or early 2026



CMS PAPER CDR-24-XYZ

DRAFT
CMS Paper

The content of this note is intended for CMS internal use and distribution only

June 11, 2025
Archive Hash: none
Archive Date: none

CMS Offline Software and Computing for HL-LHC

Conceptual Design Report

The CMS Collaboration

Abstract

This Conceptual Design Report (CDR) for Phase-2 CMS Offline Software and Computing (O&C) outlines the plan to enable the physics program of the experiment during Phase-2, due to start in 20230. While certain elements of the current software and computing infrastructure are scalable and sustainable for the foreseeable future, other aspects are not and will need to be adapted for HL-LHC, especially in light of new architectures and ways of provisioning computing resources. Modern technologies and facilities, such as heterogeneous computing, high performance computing centers, machine learning algorithms, interconnecting networks, novel memory and storage designs, open up the possibility of exploiting new capabilities and functionalities. We will step through the various areas of O&C, outlining our plans to evolve the offline software, grid middleware, and computing infrastructure, estimating in each case the positive impact of success in terms of resources, the risks of failure, etc. The management framework for the coordination of the CMS O&C Upgrade program is described; coordination and collaboration will be required not only within CMS but also with external software development communities, WLCG sites, and others. Updated projections on the computing resource needs of CMS in the 2030-2038 period will be given, both for a baseline scenario and scenarios that consider the likely outcome of the various R&D activities. A high-level version of a Technical Roadmap for the 2025-20230 preparation period is presented, and includes a timeline for milestones associated with the main infrastructure and software elements of the CMS Phase-2 computing system.

This box is only visible in draft mode. Please make sure the values below make sense.

PDFAuthor:	D. Elvira, F. Ferri, et al.
PDFTitle:	CMS Offline Software and Computing for HL-LHC CDR
PDFSubject:	CMS
PDFKeywords:	CMS, computing

Please also verify that the abstract does not use any user defined symbols

Computing resource estimates and projections - Disk

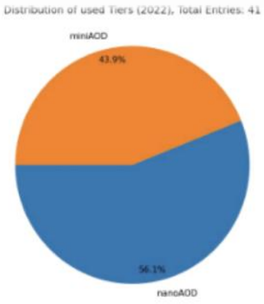
Resource: Disk Nominal ≡ baseline scenario (includes 100% prompt reco)

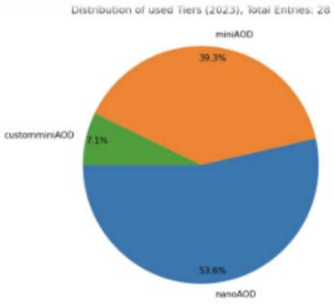
Updated annual resource increase estimates from WLCG: 10±5% (CPU), 5±5% (Disk), 10±5% (Tape)

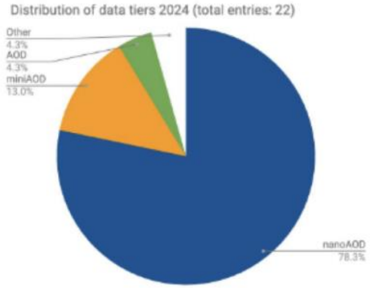
Wider adoption of nanoAODs

AOD: ≈ 500 kb/ev → miniAOD ≈ 50 kb/ev → nanoAOD ≈ 1-2 kb/ev

Data Formats Size (kB/evt.)	PU=62	PU=140	PU=200
RAW	1 200	4 300	5 900
RECO	4 300	14 000	20 000
AOD(SIM)	565 (605)	1 400	2 000
MiniAOD(SIM)	64 (86)	180	250
NanoAOD(SIM)	1.5 (1.5)	4	4
HLTScout	12	17	25
L1Scout	360		







Majority of analyses now using nanoAOD → important milestone for Phase-2 preparation

Computing resource estimates and projections - Tape

Resource: Tape Nominal ≡ baseline scenario (includes 100% prompt reco)

gap,

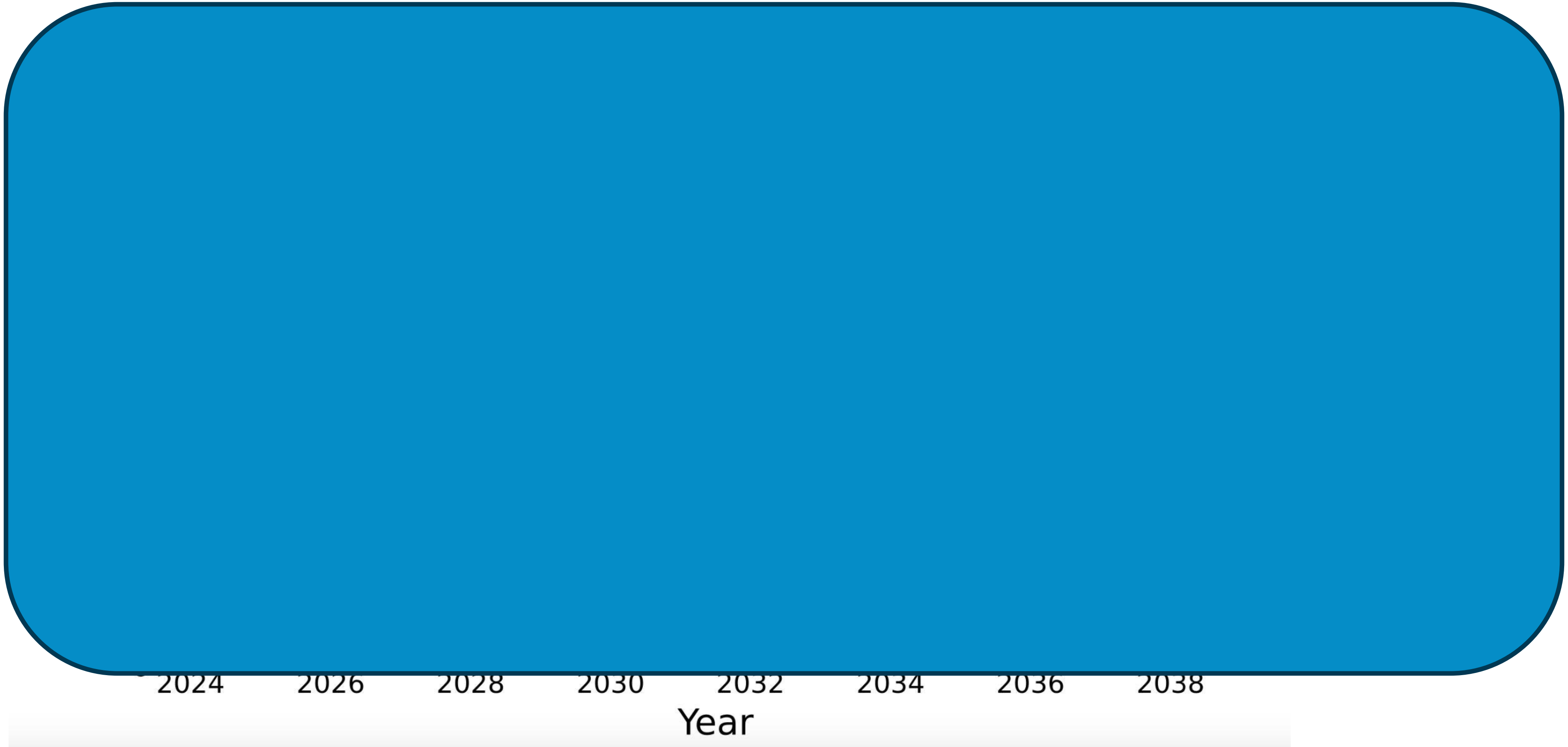
Updated annual resource increase estimates from WLCG: 10±5% (CPU), 5±5% (Disk), 10±5% (Tape)

Computing resource estimates and projections - CPU

Resource: CPU Nominal ≡ baseline scenario (includes 100% prompt reco)

Evaluate GPU needs under H, M/M', L scenarios & **CPU only vs. CPU+GPU** cost optimization

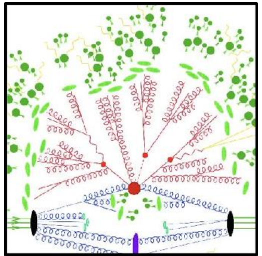
Computing resource estimates and projections - CPU



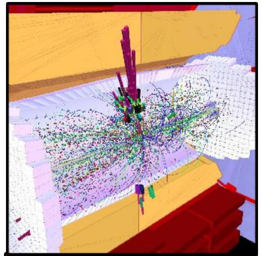
CMSSW with Heterogeneous Architectures

Real Data

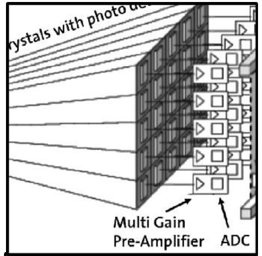
Monte Carlo



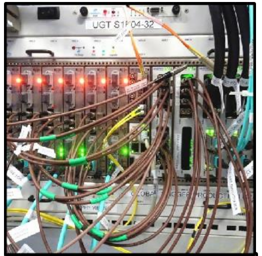
GEN



SIM

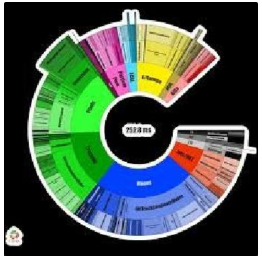


DIGI



L1T

DIGI2RAW

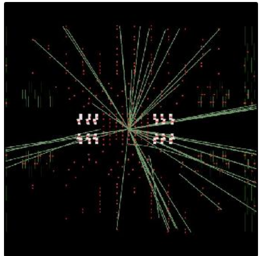


HLT

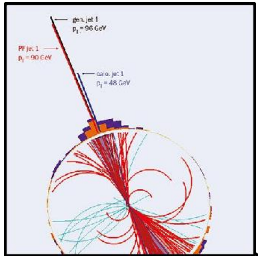
4GPU

AdePT/Celeritas
See recent HSF seminar

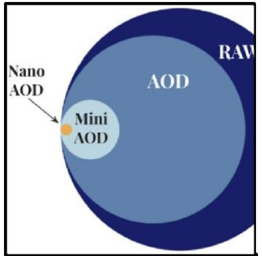
FlashSim (CMS-CR-2025-027)



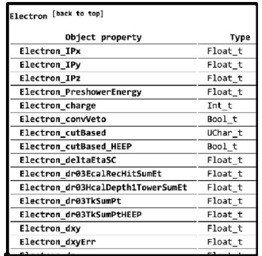
RAW2DIGI



RECO



MINIAOD



Electron [back to top]

Object property	Type
Electron_IPx	Float_t
Electron_IPy	Float_t
Electron_IPz	Float_t
Electron_PreshowerEnergy	Float_t
Electron_charge	Int_t
Electron_convVeto	Bool_t
Electron_cutBased	UChar_t
Electron_cutBased_HEEP	Bool_t
Electron_deltaEtaSC	Float_t
Electron_dr03EcalRecHitSumEt	Float_t
Electron_dr03HcalDepth1TowerSumEt	Float_t
Electron_dr03TkSumPt	Float_t
Electron_dr03TkSumPtHEEP	Float_t
Electron_dxy	Float_t
Electron_dxyErr	Float_t

NANOAOD

SONIC

Performance depend on Hardware. GPU could provide 10x more event rate.

CMSSW with Heterogeneous Architectures

Currently, **CMSSW** is being built for ARM and x86_64 architectures. Heterogeneous support is at the module level. GPU offloading development is led by TSG, targeting Run-3 HLT, and subsequently expanding to offline reconstruction

- Vertex reconstruction
- Line Segment Tracking (LST)
- The Iterative CLustering (TICL)
- PF reconstruction
- ECAL/HCAL local reconstructions
- Electron Seeding

Event Generation

Most of CMS MC generators use Matrix Element (ME) with general purpose generators for showering. Major LO backgrounds, O(10) B events, such as DY or TTbar uses MG5 to deal with hard jets up to 4 jets, then use pythia for showering, then MLM matching.

Step 1- Gridpack production

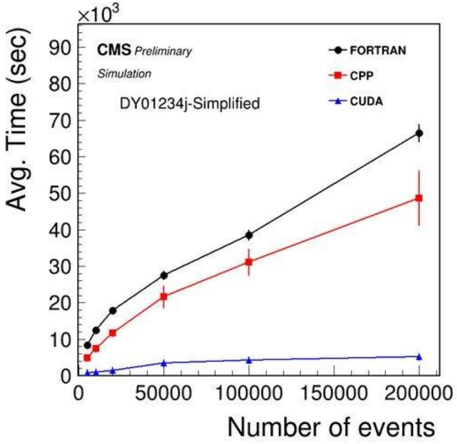
Possible to achieve speed-up using CUDA in subprocess level

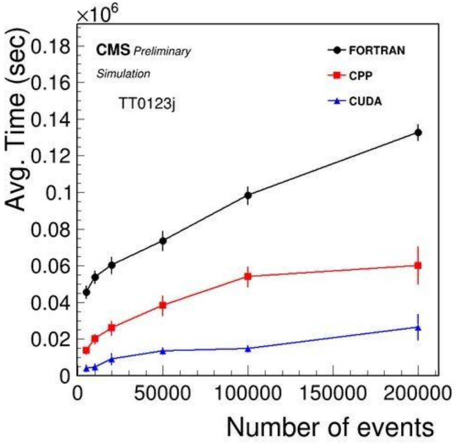
Production Time			
Process	FORTRAN	CPP-AVX2	CUDA
DY+0j	7m	6m	5m
DY+1j	10m	10m	12m
DY+2j	1h 12m	1h 10m	51m
DY+3j	22h 40m	9h 4m	4h 18m
DY+4j (Simplified)	440h 46m	141h 20m	9h 17m
DY+01234j (Simplified)	424h 36m	133h 38m	9h 32m
TT+0j	6m	7m	5m
TT+1j	11m	11m	7m
TT+2j	1h 15m	38m	22m
TT+3j	262h 11m	79h 19m	3h 4m
TT+0123j	253h 36m	155h 28m	3h 9m

Step 2 - Production

Promising results with GPU

[CMS-DP-2024-086]

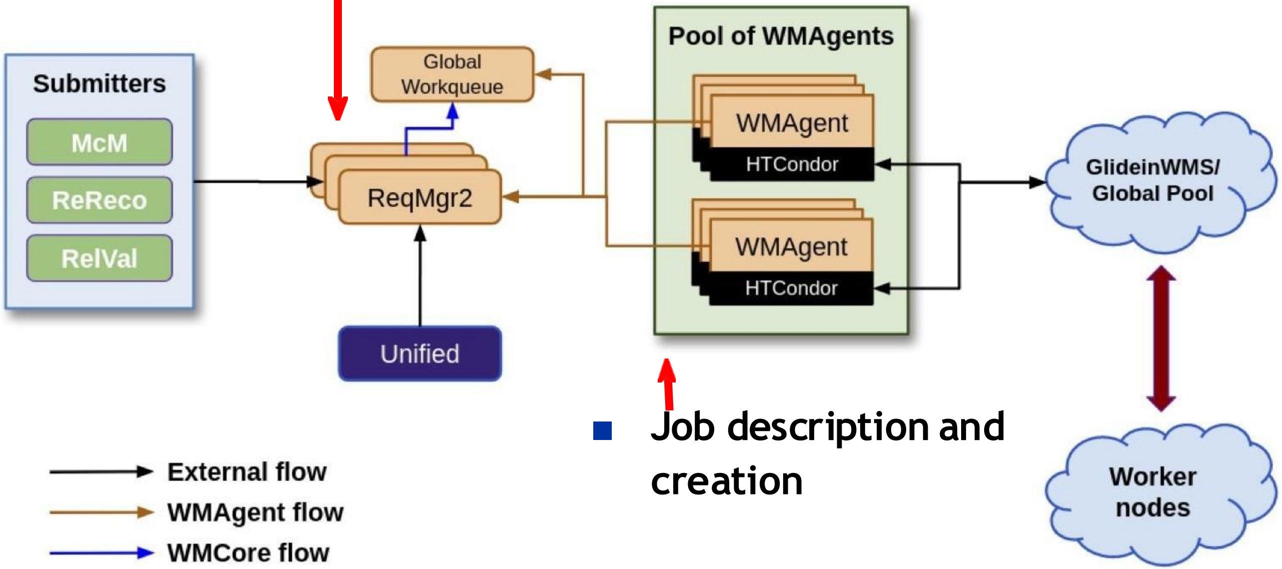




Remaining question

- For GPU: How to organize 2 steps, MG + Pythia, efficiently. Currently, Pythia8 supports multi-thread but not GPU.

- **Workflow description:**
  - **RequiresGPU**: forbidden (default), required, optional
  - **GPUParams**: Memory, CUDA runtimes, CUDA capabilities (Link), ... etc. Use of AMD and Intel GPUs not yet commissioned.



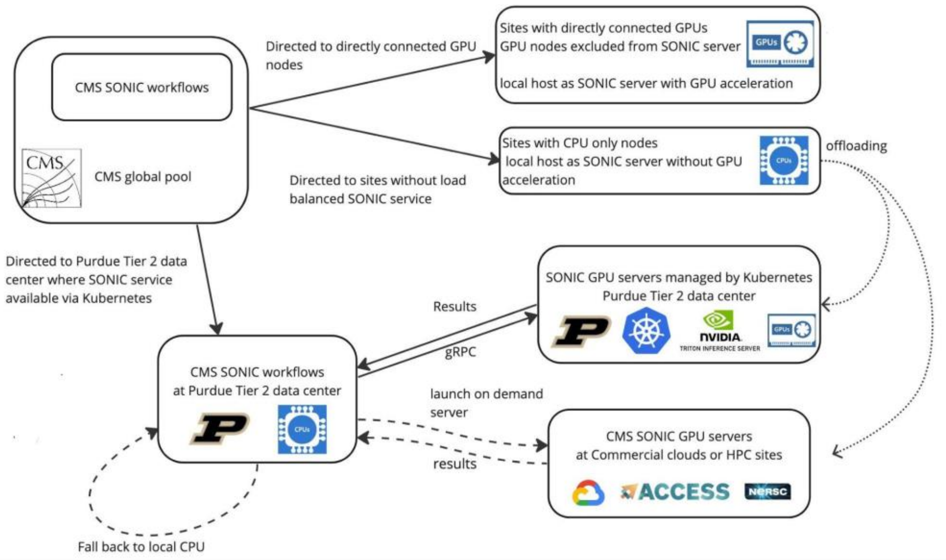
- **Job description and creation**

- **Define resource description in the global pool**
- **Match job requirements with machine attributes**

Until now, we have only discussed cases where the coprocessor is directly connected to the CPU. But what happens when the coprocessor resources are located on a separate machine?

SONIC

The idea of **SONIC (Services for Optimized Network Inference on Co-processors)** is that the inference part of the code can be sent to a remote co-processor over the networks and the results obtained (a)synchronously. Not every CPU needs a GPU sitting next to it. SONIC employs Triton Servers to manage and serve remote inference requests.



- SONIC separates workflows into "client" and "server" components that communicate via gRPC.
- In CMS, "clients" are CMSSW producers modified to asynchronously off-load inferences to servers and acquire results.
- "Server" can be a simple CPU- or GPU-powered NVIDIA Triton Inference Server, or a more complex system with load balancing over multiple GPUs and other functionalities, such as SuperSONIC.
- Models are loaded from CVMFS
- Server address is configured at site level and can be discovered by CMS production jobs automatically. It can be local or remote (e.g. HPC center).

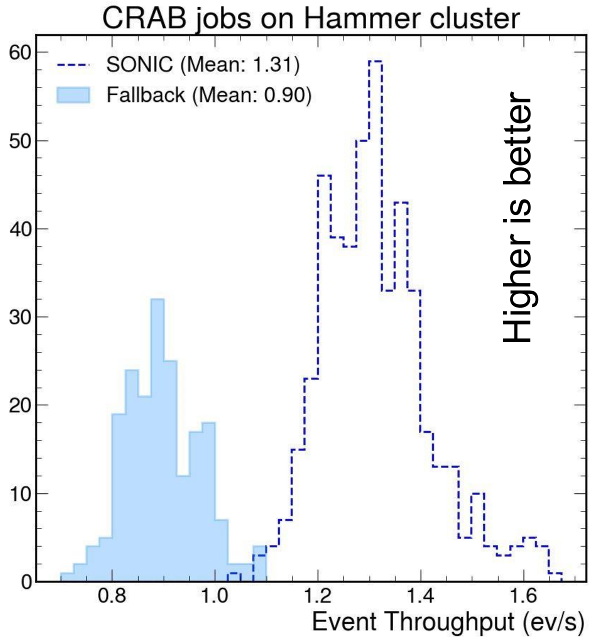
SONIC

SONIC large scale tests at Purdue Tier-2 (2024):

- Realistic workflow: CMS Run 2 MiniAOD with multiple ML models off-loade to SONIC; 1000 jobs submitted in batches.
- Sustained efficient load balancing over 9 GPUs for the duration of the test, ≃35% throughput improvement over local CPU.

Recent Developments:

- SuperSONIC: server infrastructure packaged for portability to k8s-enabled CMS sites and HPC centers. Includes improved load balancing, autoscaling, rate limiting, monitoring.
- Port more models to Triton Inference Server and keep pace with CMSSW development:
  - Machine learning Particle Flow reconstruction (MLPF).
  - Unified Particle Transformer, a Jet flavor tagging algorithm.



Heterogeneous resources : HPCs

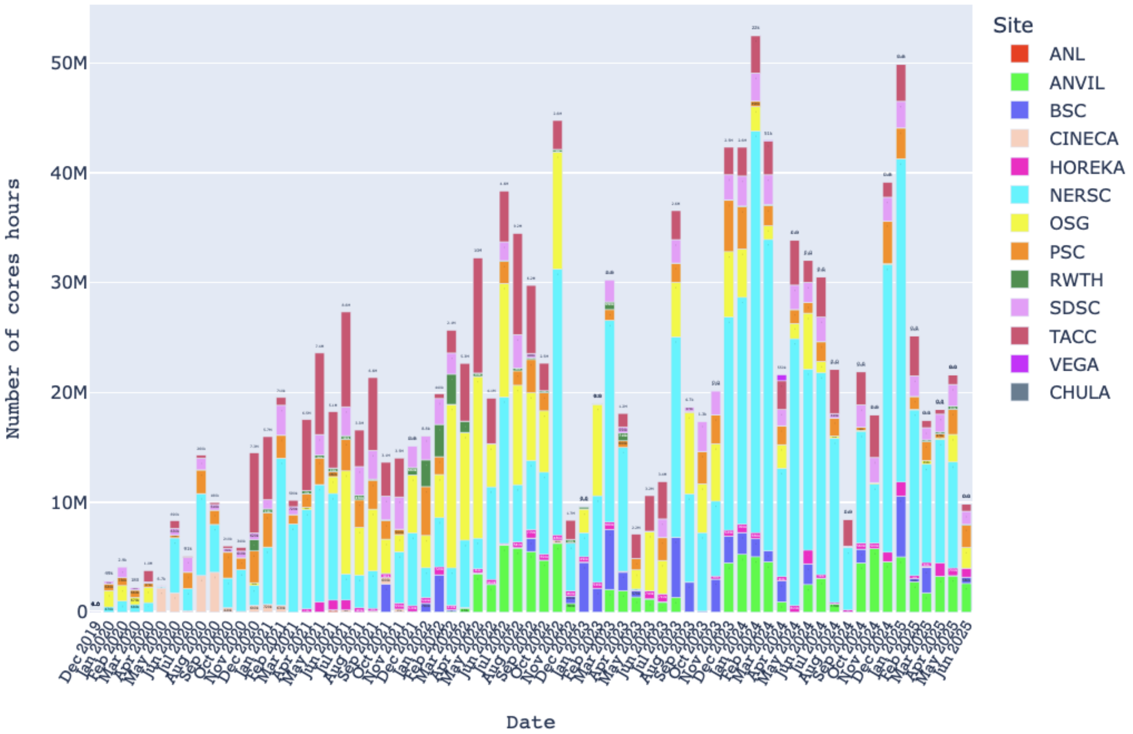
- Several HPC machines providing GPUs have been integrated into CMS system allowing technical validation of the computing system
  - Various US machines, CINECA Italy, HoreKa in Germany.
- Recently Vega, Slovenian EuroHPC, has been successfully exploited in order to contribute to the HLT Software Validation
  - Dedicated CMS grant submitted through EuroHPC
  - A fruitful synergy between Trigger, Software and Computing areas.
- Future opportunities foreseen in the context of the 1st EuroHPC AI_Factories initiative.
  - Our interest is to collaborate with funding agency and EuroHPC to make these resources available to CMS
- In addition to what we might exploit from the future Frontera upgrade in US.

HoreKa, Germany

Vega, Slovenia

HPC Sites

- Taking into account both opportunistic resources and HPC-like resources at WLCG sites (e.g. HOREKA at KIT).
- Continuously growing. Biggest part comes from US sites (NERSC on top).



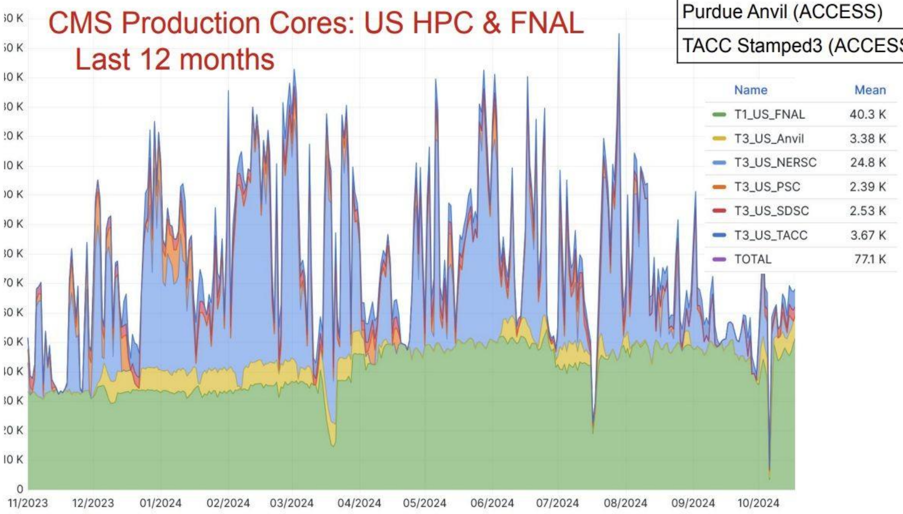
HPC Sites

- Taking into account both opportunistic resources and HPC-like resources at WLCG sites (e.g. HOREKA at KIT).
- Continously growing.

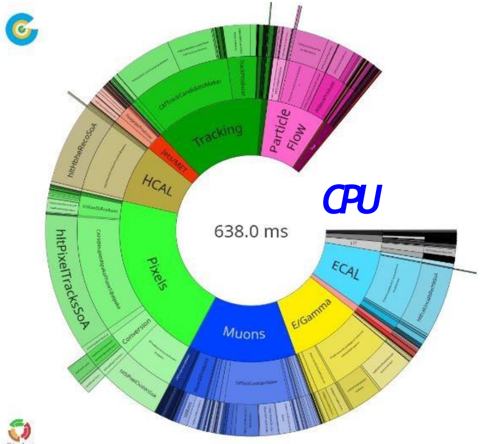
- Comparable to FNAL T1 in scale (but not continuously available/used, utilization fluctuates)

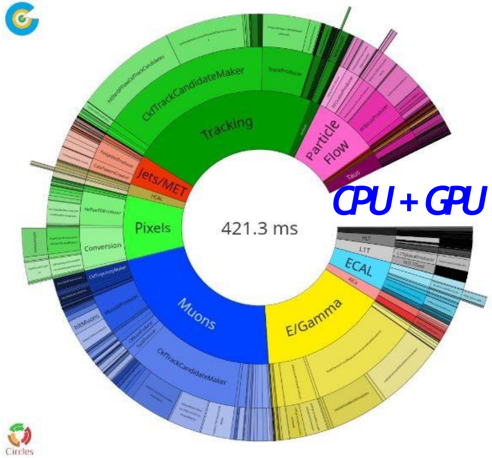
HPC Resource	Allocation (core hours)	Allocation Period	%Used*
NERSC Perlmutter (DOE)	337M CPU	Jan 2024 - Jan 2025	74%
TACC Frontera (NSF)	36M	Jun 2024 - May 2025	38%
PSC Bridges-2 (ACCESS)	23M	Oct 2024 - Sep 2025	2%
SDSC Expanse (ACCESS)	23M	Oct 2024 - Sep 2025	8%
Purdue Anvil (ACCESS)	23M	Oct 2024 - Sep 2025	15%
TACC Stamped3 (ACCESS)	1M	Oct 2024 - Sep 2025	0%

Currently Active Allocations



Online Heterogeneous Reconstruction



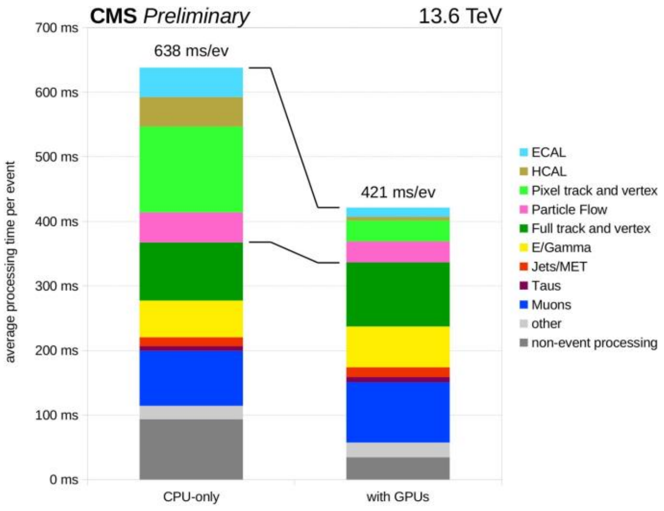


Alpaka (Abstraction Library for Parallel Kernel Acceleration)

- Alpaka reconstruction runs on CPUs and GPUs with near-identical results, validated on x86-64 and NVIDIA GPUs for HLT.

Online Reconstruction (High-Level Trigger): Currently, CMS can offload to GPUs about 35% of the online reconstruction time [Ref]:

- the ECAL unpacking and local reconstruction
- the HCAL local reconstruction and Particle Flow clustering
- the Pixel unpacking, local reconstruction, track reconstruction, and vertex reconstruction



Example of ongoing development:

- Electron seeding: ~15% of overall reconstruction time @ HLT, and ~90% of the e/gamma reconstruction time spent on electron seeding [Ref]

Alpaka Framework

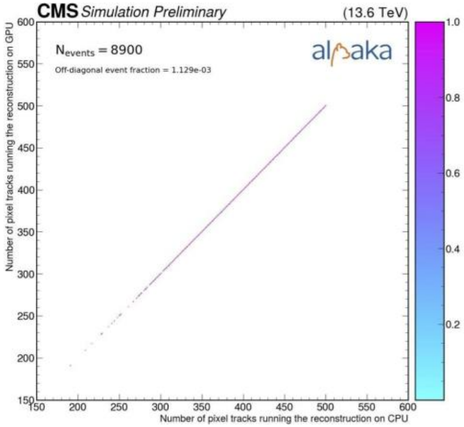
Supported GPUs in CMSSW

- NVIDIA: Production grade and validated.
- AMD: Production grade, in validation. Running CMSSW CI tests on LUMI.
- Intel: Support is not complete for CMSSW. Integration with Intel oneAPI has not been tried yet.

Accelerator Back-end	Lib/API	Devices	Execution strategy grid-blocks	Execution strategy block-threads
Serial	n/a	Host CPU (single core)	sequential	sequential (only 1 thread per block)
OpenMP 2.0+ blocks	OpenMP 2.0+	Host CPU (multi core)	parallel (preemptive multitasking)	sequential (only 1 thread per block)
OpenMP 2.0+ threads	OpenMP 2.0+	Host CPU (multi core)	sequential	parallel (preemptive multitasking)
std::thread	std::thread	Host CPU (multi core)	sequential	parallel (preemptive multitasking)
TBB	TBB 2.2+	Host CPU (multi core)	parallel (preemptive multitasking)	sequential (only 1 thread per block)
CUDA	CUDA 12.0+	NVIDIA GPUs	parallel (undefined)	parallel (lock-step within warps)
HIP(clang)	HIP 6.0+	AMD GPUs	parallel (undefined)	parallel (lock-step within warps)
SYCL(oneAPI)	oneAPI 2024.2+	CPUs, Intel GPUs and FPGAs	parallel (undefined)	parallel (lock-step within warps)

Effort to validate results

- Every module requires technical validation, where results should be numerically similar but may differ due to factors such as hardware architecture, the order of floating-point operations (e.g., in parallelized algorithms), or whether Fused Multiply-Add (FMA) instructions are used versus separate multiply and add operations.
- Using Alpaka-based modules that share nearly identical code and common data formats significantly simplifies the comparison process.



HEPSpec-like score for GPU is in development

- Currently, using CMS (HLT) workflows

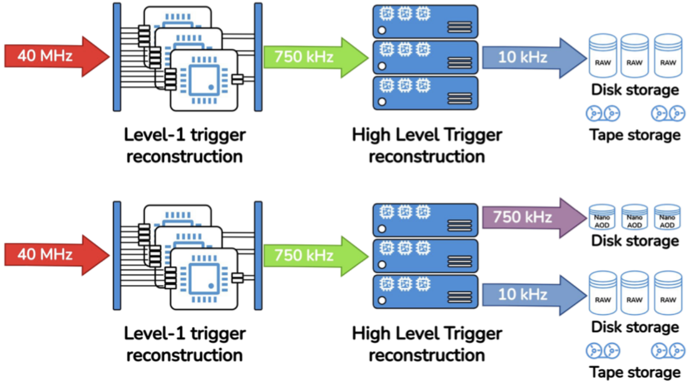
NGT

- https://nextgentriggers.web.cern.ch

Task 3.1.1: R³ Faster Reconstruction

- The successful **Patatrack experience** in CMS has shown that it is possible to **improve the physics quality and reconstruction throughput of selected physics objects (pixel tracks) by leveraging heterogeneous architectures**
- This required ~4 years of development to:
  - Study the performance of the current algorithm and identify bottlenecks
  - **Rethink the algorithms** and **data structures** targeting heterogeneous architectures
  - Develop, integrate and validate the results in CMSSW
  - Propagate the new objects to the rest of the reconstruction
- The R³ project will use a similar approach to redesign the most important physics objects:
  - Muons
  - Electrons and photons
  - Taus
  - Jets, MET and Particle Flow Global Event interpretation
- Perform offline-like full event reconstruction,

  in addition to the traditional event selection



«Heterogeneous» resources : ARM

- Successfully integrated ARM resources available at several sites.
  - In production: CNAF, available through the Global Pool
  - Testing: KIT, CERN, ScotGrid_GLA, Ookami
- Currently, we are working towards the integration of Deucalion, ARM based EuroHPC Supercomputer located in Portugal:
  - Access granted via Openlab at CERN
  - Activity still at early stage, from technical perspectives the system is not ideal (still not providing access to CVMFS CERN repos) but "promising".
  - Integration will be based on existing Submission Infrastructure tools via the Global Pool
- On the US side, the next large NSF HPC, the Frontera upgrade, it's currently in early construction.
  - Although not 100% confirmed, It's likely going to be partitioned into NVIDIA ARM CPU and NVIDIA ARM CPU + GPU

```
CMSHTPC_T1_DE_KIT_htcondor-ce-1-kit_arm
CMSHTPC_T1_DE_KIT_htcondor-ce-2-kit_arm
CMSHTPC_T1_DE_KIT_htcondor-ce-3-kit_arm
CMSHTPC_T1_DE_KIT_htcondor-ce-4-kit_arm
CMSHTPC_T1_IT_CNAF_CINECA_Marconi100_arm
CMSHTPC_T1_IT_CNAF_Deuclion_arm
CMSHTPC_T1_IT_CNAF_condor_ce02_arm
CMSHTPC_T1_IT_CNAF_condor_ce03_arm
CMSHTPC_T1_IT_CNAF_condor_ce04_arm
CMSHTPC_T1_IT_CNAF_condor_ce05_arm
CMSHTPC_T1_IT_CNAF_condor_ce06_arm
CMSHTPC_T2_CH_CERN_ce511_arm
CMSHTPC_T2_CH_CERN_ce512_arm
CMSHTPC_T3_UK_ScotGrid_GLA_ce_arm
CMSHTPC_T3_US_Ookami
```

Deucalion, part of EuroHPC

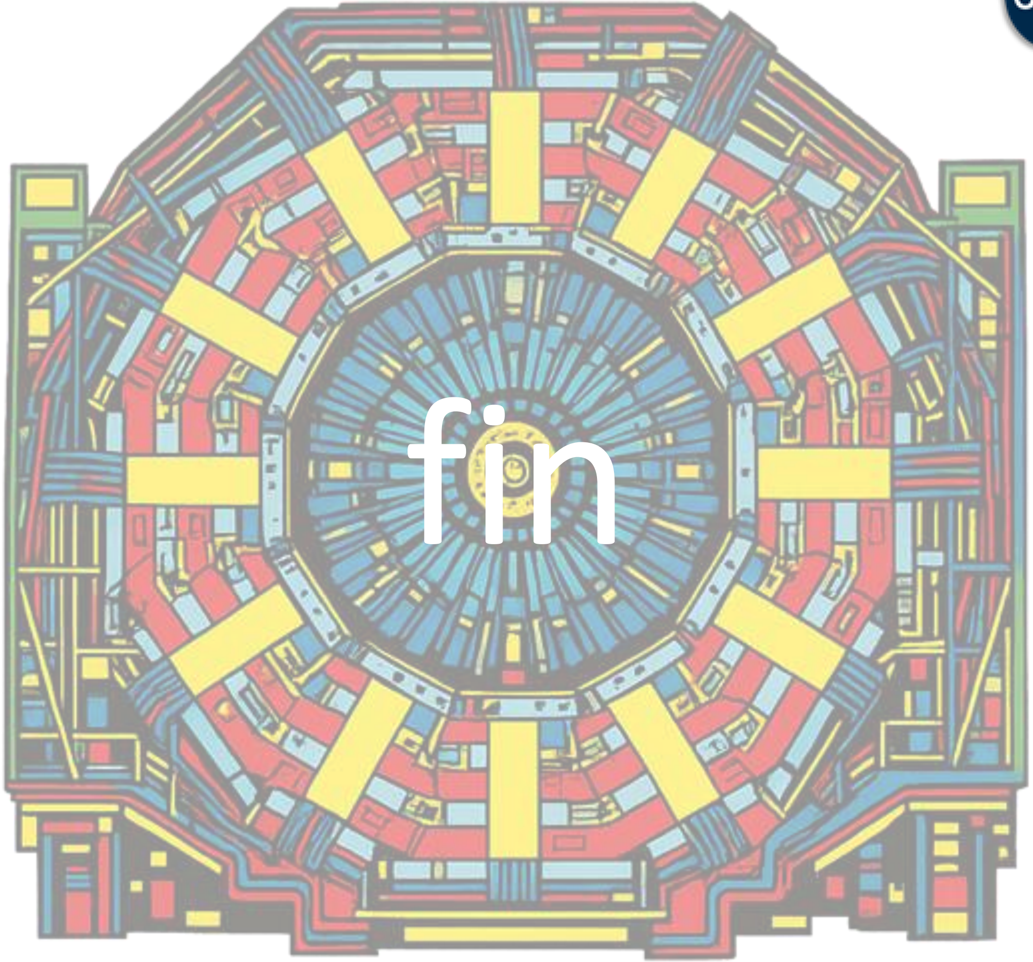
Summary

- Run 3 is now in full swing, surpassing Run 2. Still ~1.5y to go to get past the final goal of doubling Run 2.

- CMS is continuing to utilize computing resources intensively but efficiently (in France and elsewhere).

- Phase-2 preparations continue to ramp up:

    - The CDR is converging later this year. Target is to have it fully public in early 2026.
    - The preliminary numbers show that maybe we were a bit optimistic.
    - **BUT** a plan of action is ready. Its foundations rely on a shift towards heterogeneous architectures.


fin