WLCG Workshop on Heterogeneous Architectures

(including recent progress and evolution)

O. Smirnova (Lund), A. Di Girolamo (CERN), and J. Letts (UCSD),
on behalf of the WLCG Technical Coordination Board

HEPiX Spring Workshop - TechWatch Session - April 22, 2026

WLCG Workshop on Heterogeneous Architectures

WLCG held a three-day Workshop on Heterogeneous Architectures in December 3-5, 2025 at CERN.

- Indico: https://indico.cern.ch/event/1526077/
- There are extensive live notes with an executive summary at the end, as well as a report for the WLCG MB distilled in the weeks after the Workshop.
- This presentation will outline the major takeaways as detailed in the report to the MB as well as future plans and directions.
- The outcomes of the Workshop are an important input for the drafting of Chapter 8 of the WLCG Technical Roadmap on “New Architectures and Infrastructures”.
- There will likely be a follow-up session in the upcoming HSF-WLCG Workshop in Bologna in November 2026. This follow-up was considered important by many attendees in December.

Main Takeaways from the Workshop

Main Takeaways: Experiments' Perspectives

- GPU acceleration is a vital R&D area in all four LHC experiments in many areas, including event generation, detector simulation, event reconstruction, ML, inference, etc.
- However, no experiment is ready to formally request or accept pledges of GPUs: still considered an area of active R&D at this time.
- Note that the resource request process begins about 18 months before deployments: Since Run 4 begins in 2030, the code bases, computing models, benchmarking & accounting all need to be in place *before* the summer of 2028 (which is also around the time when experiments may be planning their computing readiness challenges for HL-LHC).

*These are the "gaps" that need to be bridged over the next **two years**.*

The schedule is tight: fortunately much progress has been and continues to be made!

Main Takeaways: Sites' Perspectives

Requesting GPUs is only one side of the challenge - the other is provisioning.

Sites need to not only understand how many GPUs to provide but also:

- The various use cases: ML training and inference, physics analysis, (including "Large" physics models), and "traditional" GPU calculations.
  - These may require different infrastructures, access methods (batch vs. interactive).
  - Here the evolution of the workflow management (Chapter 5 of the WLCG Technical Roadmap is very relevant!
- CPU-to-GPU ratios
- Role of HPCs and "opportunistic" GPU resources in general.
- Total Cost of Ownership & Pricing Evolution - This has become more and more important - and complex to estimate - over the past year!

Bridging the “Gaps”

LHC Experiments' Computing Models

Scale (% of *offline* processing wallclock time) & type of the offload foreseen (including some updates since the Workshop, where known):

- ALICE: **Today 50%-60% of full offline processing** on GPU yielding 2x to 2.5x speedup. Optimistic target is 80% with full barrel tracking, aiming for 5x speedup. *N.B.* Online reconstruction almost completely ported to GPUs
- ATLAS: Foresee being able to offload 50% of full simulation, approximately equivalent to **15% of their total offline processing** requirements.
- CMS: Phase-2 Software and Computing CDR recently published with R&D drives to offload many steps of the offline workflow. Foresees **between 15%-40% offload in total** depending on the outcome of these various R&D activities (updated since the Workshop, also at the January 2026 WLCG MB#336 meeting).
- LHCb: Full simulation (most of their offline processing needs) can run on AdePT with **~2x speedup**. Great collaboration with CERN EP/SFT group.

LHC Experiments' Computing Models

- While no experiment is ready to request or accept pledges of GPUs at this time, all are planning on leveraging heterogeneous resources in Run 4.
- Experiments are making updated code available for inclusion in the HEPScore4GPU benchmark as it becomes available, including GPU-enabled code based on software such as AdePt for detector simulation.

Open Question at the Workshop: When can we quantify how much GPU capacity will be available on experiment's new or previous HLT farms and their availability for offline use?

Software Evolution

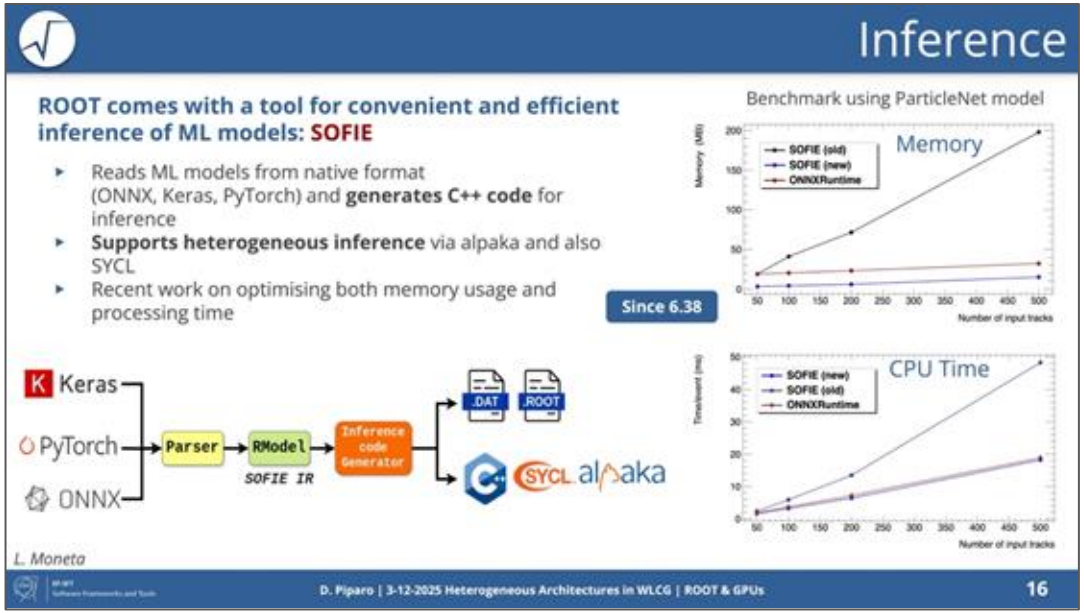
Many software packages that underlie experimental stacks have their own GPU evolution planning, including:

- Event generators: MadGraph, Pepper, others...
- Detector Simulation: AdePt and Celeritas - with close collaboration with the experiments.
- Detector Geometry: There was some discussion about addressing the scalability and compilation complexities of VecGeom and experiment geometries in general.
- ROOT: Many new features or plans for advancement in statistical interpretation (in production) ML training (in production) and inference (see next slide), and data analysis. ROOT also offers easy ways to plug-in new software technologies leveraging GPUs, e.g. future (de-)compression libraries.

Open Question at the Workshop: What are the resource needs for ML training and do they need to be provided by specialized ML training facilities or can HPC allocations, university clusters, or national platforms fulfill these needs?

Example: ML Inference with ROOT

While not a direct focus of this Workshop, analysis tools and infrastructures are also expected to leverage GPUs heavily for training and inference.



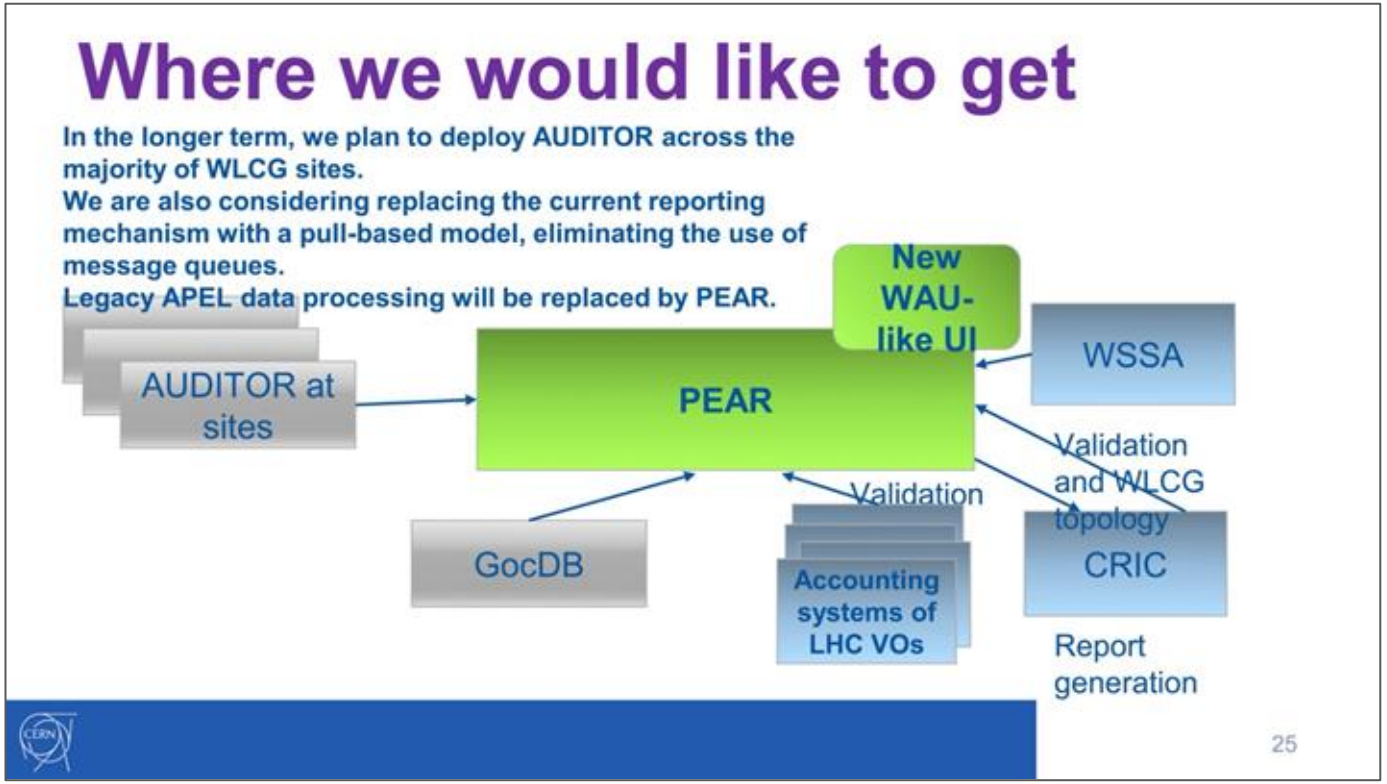
D. Piparo, WLCG Workshop on Heterogeneous Architectures

Information Systems & Accounting

The evolution of WLCG information and accounting systems (see also the next slide) was discussed in the Open Technical Forum #9 last month [Indico] as well as in the December Workshop:

- Very important for GPU discovery and advertising of heterogeneous resources
- Need to agree on a description schema.
- Inherent heterogeneity itself (GPU models, driver versions etc.) was seen as a challenge in handling job failures, compilation challenges, portability of code, etc.
- Understanding how to improve the monitoring of GPU occupancy and measure performance.
- Based on that improved monitoring, develop accounting systems of GPU performance and utilization.

Information Systems and Accounting



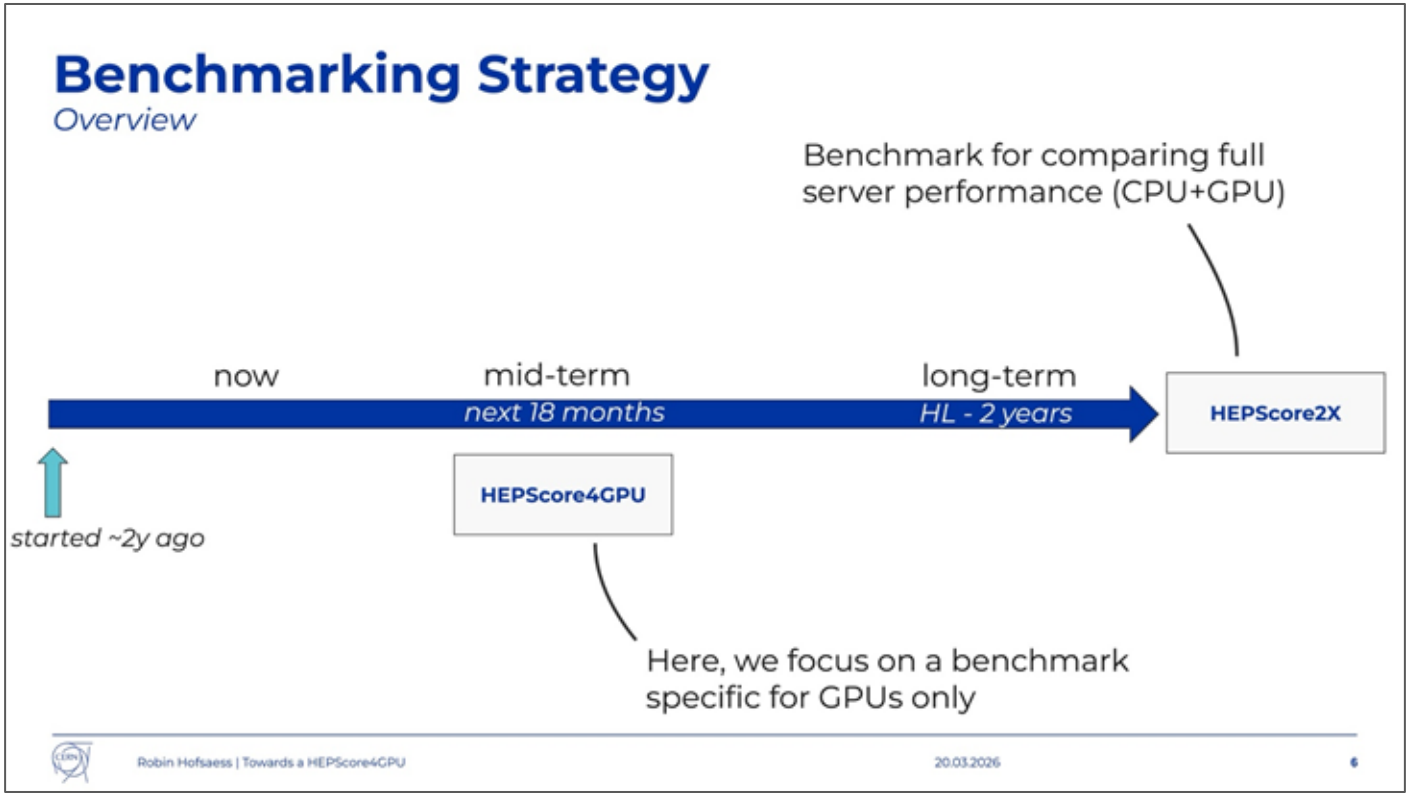
J. Andreeva et al., OTF#9

Benchmarking GPUs

GPU Benchmarking is an essential piece for quantifying pledges. Benchmarking was discussed extensively in the December Workshop, and rapid progress has taken place since then, including the including of new experiment GPU workflows and defining the benchmarking framework of HEPScore4GPU.

- There was a presentation from the HEPiX Benchmarking Working Group earlier today.
- See also this recent presentation with milestones and timelines in an Open Session on Heterogeneous Software and Computing in the CMS Offline & Computing Week, as well as other presentations on GPU resources, software frameworks, portability layers, provisioning, etc. in the Open Session.
- Measurement of the relative performance of different GPU vendors' models seems possible even now, and may help guide initial GPU deployments by Site Federations in early LS3.

Benchmarking GPUs



R, Hofsaess, D. Giordano, March 20, 2026

Workflow Management

Workflow Management (WM) systems need to evolve to co-schedule CPU and GPU workflows e.g., simulation, reconstruction, but also training and inference, handling partitionable GPUs, checkpointing, etc.

- Much work taking place evolving WM systems in the experiments, including in the major shared frameworks such as PanDA, DiracX, et al.
- Plans to “bridge the gaps” in WM will be handled in Chapter 5 of the WLCG Technical Roadmap, where areas for coordinated action will be outlined, including:
  - Improved integration of opportunistic and allocation-based resources
  - Capability-aware resource matchmaking
  - Common approaches to heterogeneous resource utilisation
  - Evolution of infrastructure interfaces
  - Cross-experiment coordination on provisioning tools and concepts
  - Support for emerging workflow paradigms

Open Question (post-Workshop) about data fragmentation e.g., what if a site has the data but not the GPUs to process it? This was not explicitly discussed in the workshop.

Education and Training

As with any “new” technology, providing education and training to better match user needs with the new infrastructures (e.g., the CERN ML infrastructure) was seen as an important pillar of the bridge.

- Documentation on how to train ML models
- Including on HPCs and other “external” resources that may have non-experiment-supported workflow submission systems

Training providers will likely be spread all throughout the WLCG and beyond.

Sharing information will be key to successfully enabling the participation of the entire community.

Cost Evolution

Cost evolution of GPUs (and computing hardware in general) was presented but not extensively discussed during the Workshop.

- The consensus was that it would be a very helpful guide to WLCG Federations to optimize multi-year funding requests and procurements to establish shared cost predictions for GPUs, which could be done in the context of the HEPiX TechWatch working group.

We also want to understand the “Total Cost of Ownership” of GPUs including power consumption. While not discussed during the Workshop, the work of both the WLCG Environmental Sustainability Forum and the HEPiX TechWatch Working Group are very important here.

- Chapter 9 of the WLCG Technical Roadmap will focus on both environmental and technical sustainability aspects of the technical evolution of WLCG sites and services.

Cost Evolution

Relative costs of GPU vs. CPU processing capacity (in HEPScore units) will be essential in making the decision ***if and how much*** to offload, once it is technically possible.

- Recent “AI Boom” has severely disrupted the market, especially for RAM and GPUs.

“It’s tough to make predictions, especially about the future.”

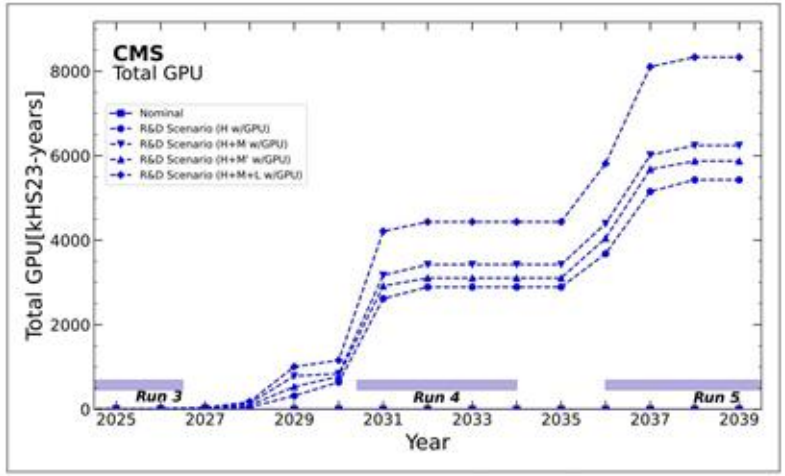


Figure 14: Estimated total GPU processing needs annually through HL-LHC Run 5. Need is expressed in terms of HS23 that has been offloaded. Blue curves correspond to annual projections of needs for the baseline scenario and separately for each of the R&D scenarios. Section 11.3.1 describes current estimates for the conversion of HS23 units into the actual needed GPUs.

CMS CDR, CERN-LHCC-2026-003

The Path Forward

GPUs are increasingly becoming part of the compute landscape and WLCG needs to be able to exploit these. A holistic approach is needed for this multifaceted problem. The main macro areas we see are:

- Benchmarking:
  - What is the HS23 equivalent of CPU-GPU systems? This is more complicated than benchmarking CPU because it will evolve over time as workflows make better use of the GPU.
- Information System/Accounting:
  - How is the information captured and reported? How dynamic is it? How detailed is it? This is more complex than with CPU.
- Provisioning (WLCG facilities, Cloud, HPC, etc):
  - How do the resources need to be presented in order for WLCG workloads to exploit?
- Online and Offline Experiments' Software and Workflows:
  - How does (or can) software need to evolve?

Our role now is to be technically ready to be able to catch the opportunities/FAs decisions that might arise/we are forced to: TCO and agreement on accepting these pledges from the Experiments are a different dimension.

WLCG is developing a community-driven Technical Roadmap (see also yesterday's WLCG OTF co-joint with HEPiX session on Facilities Evolution).

- A specific chapter is devoted to the "New Architectures and Infrastructures".
- The milestones will be then followed up in the coming years toward HL-LHC

Conclusions

We have outlined the perceived challenges in requesting, deploying, utilizating, benchmarking, and measuring the performance of heterogeneous resources (GPUs) in HL-LHC that were discussed in the December 2025 Workshop on Heterogeneous Architectures.

- Work is progressing in many areas in the experiments, WLCG, and site federations.
- Much of the work needs to be done in the next two years (or sooner!) in preparation for experimental computing challenges and the resource request scrutiny round for 2030, which begins in 2028.

LS3 is not a shutdown for software and computing, but rather the busiest period of the decade! Everything needs to converge in the next 18-24 months.

The WLCG Open Technical Forum and WLCG Workshops, as well as co-hosted events like today's, are important forums for checkpointing progress on technical evolution in WLCG. The November HSF-WLCG Workshop will likely have a follow-up session on Heterogeneous Architectures. There will also be a plenary presentation at CHEP26 in May.

Backup Slides (for information)

Administrative Matters

There are 3 relevant CERN e-groups:

- `wlcg-technical-coordinators` (to contact Ale & James)
- `wlcg-technical-coordination-board` (the membership of the TCB)
- `wlcg-open-technical-forum` (**open to self-subscription, please join!**)

and two new Indico categories:

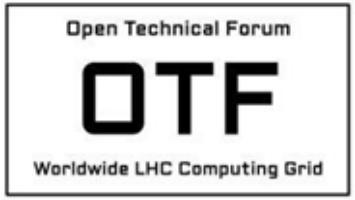
- WLCG Technical Coordination Board: https://indico.cern.ch/category/18888/
- WLCG Open Technical Forum: https://indico.cern.ch/category/18889/

The GDB e-groups will eventually be archived but for now we are still also using project-lcg-gdb to publicize OTF meetings.

e-groups to contact Technical Roadmap Chapter Facilitators

- wlcg-technical-roadmap-ch2-facility-evolution
- wlcg-technical-roadmap-ch3-data-management
- wlcg-technical-roadmap-ch4-network-management
- wlcg-technical-roadmap-ch5-workflow-management
- wlcg-technical-roadmap-ch6-security-and-aai
- wlcg-technical-roadmap-ch7-services
- wlcg-technical-roadmap-ch8-new-architectures
- wlcg-technical-roadmap-ch9-sustainability
- wlcg-technical-roadmap-facilitators-all **(all of the above)**

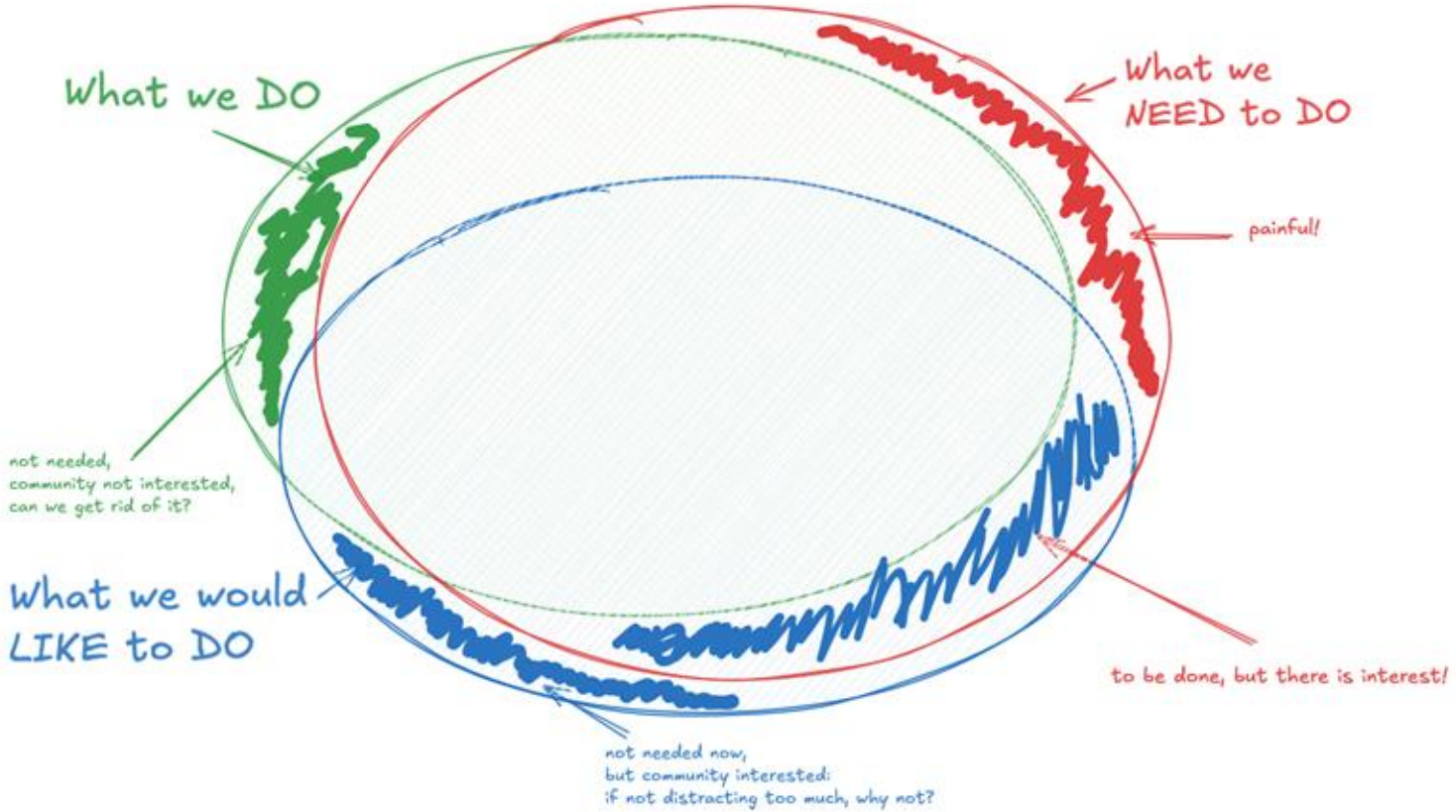
Open Technical Forum - Next Dates



- OTF#11: May 19-20th at CERN on the topic of the WLCG Technical Roadmap and a practice talk for CHEP26
- OTF#12: At the end of August (25-26th), topics to be decided
- OTF#13: September 14-18 at the XRootD and FTS workshop in Lyon (pending discussions with the organizers)

The 2026 WLCG-HSF Workshop will take place November 2nd-6th in Bologna.

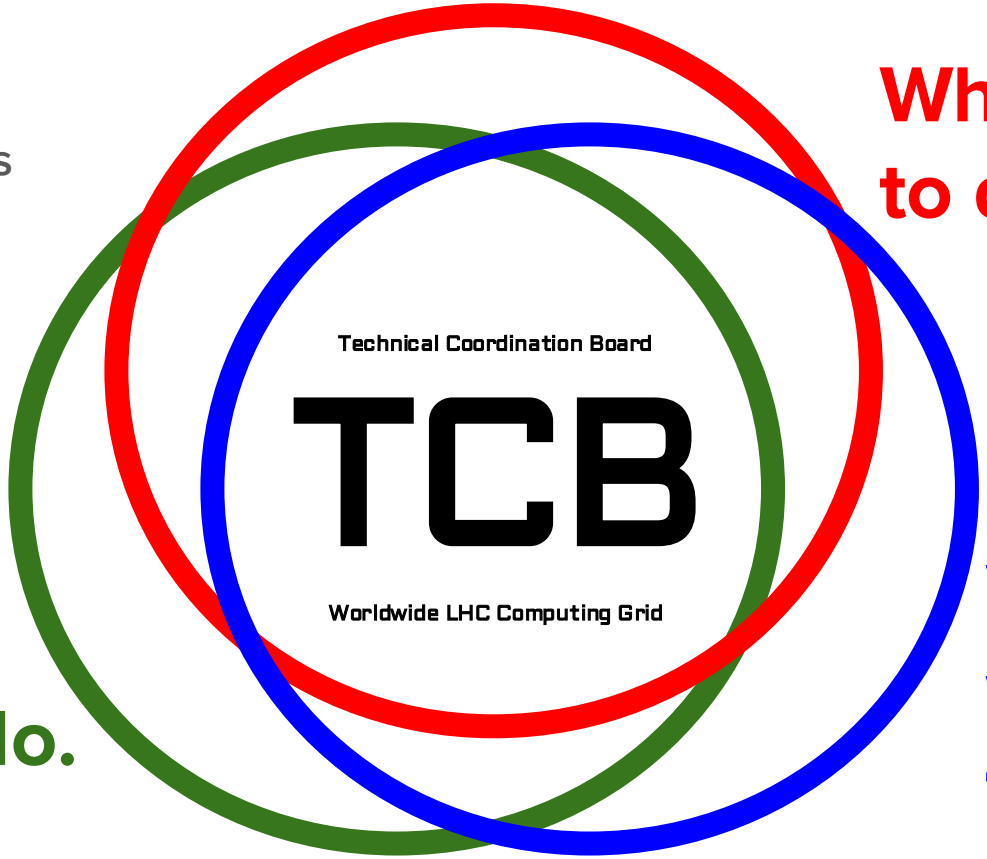
- While the agenda is still being discussed, there will likely be focus sessions on Heterogeneous Architectures (follow up to the December 2025 Workshop) and planning for DC27 (scheduled to begin in late February 2027).


What we DO
What we NEED to DO
painful!
not needed,
community not interested,
can we get rid of it?
What we would LIKE to DO
to be done, but there is interest!
not needed now,
but community interested:
if not distracting too much, why not?

Bringing Capabilities, Interests, and Needs into Focus

The TCB will aim to:

- Identify the gaps
- Develop the Technical Roadmap
- Facilitate the coordination of stakeholders to execute the Roadmap.



What we need to do.

What we would like to do.

What we do.

Key identified tasks

- Expand GPU benchmarking (HEPScore)
- Integrate more GPU-enabled codes
  - Offline Experiment Code for GPUs
  - VecGeom scalability and complexity
  - Compression/decompression
- Understanding Resource Requirements
- Run production-level workflow tests
  - Improve WM capabilities
- Education and Training
- WLCG Information System
  - Develop resource discovery standards
- Track performance & cost metrics
  - Monitoring
  - Accounting
  - Cost Estimates
  - Total Cost of Ownership