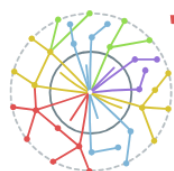


EVALUATION REPORT 2025



NextGen
Next Generation Triggers



Contact information:

CERN

Espl. des Particules 1/1211, 23 Genève

Next Generation Triggers:

<https://nextgentriggers.web.cern.ch/>

Project Coordinator:

Alberto Di Meglio

Email: alberto.di.meglio@cern.ch

Table of Contents

Executive Summary	1
Evaluation Objective(s)	1
Evaluation Methodology	1
NextGen Annual Milestones Status	2
Summary of Main Scientific, Technical and Outreach Achievements in 2025	4
Project Management and Communications	7
<i>General Evaluation</i>	<i>7</i>
<i>M2.0.1 Project management, risk management, activities and resources report</i>	<i>9</i>
Work Package 1: Infrastructure, Algorithms and Theory	11
<i>General Evaluation</i>	<i>11</i>
<i>M2.1.1 Purchase of hardware and services and commissioning completed for on-premise and cloud resources</i>	<i>11</i>
<i>M2.1.2 hls4ml software release 1 with open-access documentation</i>	<i>12</i>
<i>M2.1.3 Define and document new-physics scenarios to evaluate trigger performance. Develop and deploy quantum circuit simulations for large systems, up to $O(100)$ qubits, using tensor networks and state-vector.</i>	<i>13</i>
<i>M2.1.4 Workshop at CERN for discussing the status and plans for all of the sub-projects in WP1.7</i>	<i>14</i>
Work Package 2: Enhancing the ATLAS Trigger and Data Acquisition	17
<i>General Evaluation</i>	<i>17</i>
<i>M2.2.1: First integration of ML in L0 Global trigger for commissioning and preparation of further improvements</i>	<i>21</i>
<i>M2.2.2: Prototype GNN tracking algorithm based on ACTS</i>	<i>23</i>
<i>M2.2.3: Prototype ML and ACTS based muon reconstruction algorithm</i>	<i>24</i>
Work Package 3: Rethinking the CMS Real Time Data Processing	28
<i>General Evaluation</i>	<i>28</i>
<i>M2.3.1 A validation suite that accurately measures the performance of the R^3 reconstruction for key physics objects and representative physics signals under realistic data taking conditions is developed and integrated in CMSSW</i>	<i>30</i>
<i>M2.3.2 Creation of a small-scale prototype that buffers 30% of the RAW data for about 8 hours, derives improved calibrations, and uses them for the online reconstruction of the “HLT scouting” data stream</i>	<i>31</i>
<i>M2.3.3 First prototype Phase-2 L1 Scouting system demonstrating data acquisition and real-time physics analysis in simple final states with present technologies (i.e. Virtex Ultrascale+ HBM, 100 GbE networking, current CPU/GPUs), and first conceptual design of next generation ATCA data acquisition board for L1 Scouting (Versal HBM, 400 GbE). Results documented as CMS public notes and conference talks/proceedings</i>	<i>32</i>
<i>M2.3.4 Report on operational experience and achieved physics performance for the continuous training and deployment of ML algorithms in the L1 Trigger on Run 3</i>	<i>33</i>
Work Package 4: Education Programmes and Outreach	35
<i>General Evaluation</i>	<i>35</i>

<i>M2.4.1 2nd NextGen Triggers Project Workshop. Report on exchange and outreach activities</i>	<i>36</i>
<i>M2.4.2 Skills gaps analysis done. Report on first year of the STEAM Programme activities</i>	<i>38</i>
Conclusions and Further Recommendations	40
Appendix 1	42
Milestones for the Coming Year 2026	43

Executive Summary

This “Evaluation Report” (from now on the EV) is part of the material to be submitted to Hillspire in compliance with the agreed yearly assessment process as defined in the [Grant Agreement](#) between CERN and the Fund. It complements the “Expenditure Responsibility Report” with a brief narrative describing the achievements towards achieving the purpose of the project.

Purpose of the evaluation

This EV is produced by the NextGen Project Management Committee and delivered to the Fund through Hillspire according to the schedule agreed in the Grant Agreement, specifically by January 30th, 2026. The EV is provided to document progress, identify achievements, justify the use of resources, and provide brief recommendations and description of next steps for the following reporting period.

Scope of the evaluation

This project report covers the activities performed during the calendar year 2025 and the status of milestones **M2.0.1 through M2.4.2** as described in the Annex 3 of the Grant Agreement.

Evaluation Objective(s)

The objectives of the evaluation are to ascertain the completion or lack thereof of the milestones defined in the Grant Agreement to define the budget allocated to the project during the reporting period. For each milestone, the report provides a brief narrative description and pointers to the agreed proofs of execution. The description provides a means to the Fund to assess the extent to which the milestone has been achieved, the relevance and coherence to the original purpose, the validity of any deviation and the reasons for such deviations, and any proposed corrective actions or improvement to the main project narrative provided as part of the Grant Agreement.

Evaluation Methodology

The EV has been produced by the NextGen Project Management Committee over a period of three months (November 2025 to January 2026). It is composed of information collected from the NextGen project Task Leaders and technical experts. It has been validated and endorsed by the NextGen Steering Board in January 2026. The definition and mandate of the governance of the project is part of the Project Management Plan, which is a deliverable of Year 1 and updated in December 2025 with any relevant change in governance and structure.

NextGen Annual Milestones Status

Year	Code	Milestones	Type	Status
2	M2.0.1	Project management, risk management, activities and resources report	Report	Complete
2	M2.1.1	Purchase of hardware and services and commissioning completed for on-premise and cloud resources.	Report	Complete
2	M2.1.2	hls4ml software release 1 with open-access documentation	Software, Documentation	Complete
2	M2.1.3	Define and document new-physics scenarios to evaluate trigger performance. Develop and deploy quantum circuit simulations for large systems, up to O(100) qubits, using tensor networks and state-vector	Reports, Event	Complete
2	M2.1.4	Workshop at CERN for discussing the status and plans for all of the sub-projects in WP1.7.	Event, Report	Complete
2	M2.2.1	First integration of ML in L0 Global trigger for commissioning and preparation of further improvements.	Software Report	Complete
2	M2.2.2	Prototype GNN tracking algorithm based on ACTS	Software, Demo	Complete
2	M2.2.3	Prototype ML and ACTS based muon reconstruction algorithm	Software, Demo	Complete
2	M2.3.1	A validation suite that accurately measures the performance of the R^3 reconstruction for key physics objects and representative physics signals under realistic data taking conditions is developed and integrated in CMSSW.	Software	Complete
2	M2.3.2	Creation of a small-scale prototype that buffers 30% of the RAW data for about 8 hours, derives improved calibrations, and uses them for the online reconstruction of the “HLT scouting” data stream.	Demo	Complete

2	M2.3.3	First prototype Phase-2 L1 Scouting system demonstrating data acquisition and real-time physics analysis in simple final states with present technologies (i.e. Virtex Ultrascale+ HBM, 100 GbE networking, current CPU/GPUs), and first conceptual design of next generation ATCA data acquisition board for L1 Scouting (Versal HBM, 400 GbE). Results documented as CMS public notes and conference talks/proceedings.	Demo	Complete
2	M2.3.4	Report on operational experience and achieved physics performance for the continuous training and deployment of ML algorithms in the L1 Trigger on Run 3.	Report1 , Report2	Complete
2	M2.4.1	2nd NextGen Triggers Project Workshop. Report on exchange and outreach activities	Event , Report	Complete
2	M2.4.2	Skills gaps analysis done. Report on first year of the STEAM Programme activities	Skills Gap Analysis Report	Complete

Summary of Main Scientific, Technical and Outreach Achievements in 2025

Work Package 1 – Infrastructure, Algorithms and Theory

- **Established a shared, production-grade R&D backbone** for the project, including a fully operational ML/HPC cluster with advanced MLOps capabilities, enabling cross-WP algorithm development and integration.
- **Delivered the hls4ml v1.2.0 major release**, elevating it to a full compiler stack with bit-exact quantization, transformer support, ASIC workflows, and multi-vendor FPGA backends, now widely adopted beyond NGT.
- **Defined and initiated key new-physics benchmark scenarios** (flavour physics, dark sectors, long-lived particles), aligning theory and experiment to quantify the physics reach of next-generation triggers.

Work Package 2 – Enhancing the ATLAS Trigger and DAQ

- **Achieved the first ML integration in the ATLAS Level-0 Global Trigger**, including an automated ML-to-firmware pipeline and a prototype AI-enabled Global Common Module, marking a paradigm shift in trigger development.
- **Delivered architectural changes adopted as ATLAS Phase-2 baselines**, notably the unified readout architecture, reducing system complexity and server count by a factor of ~3.
- **Demonstrated mature ACTS- and ML-based tracking and muon reconstruction prototypes**, with validated performance gains on CPU, GPU, and FPGA, informing 2026 technology-choice decisions.

Work Package 3 – Rethinking CMS Real-Time Data Processing

- **Delivered a fully integrated Phase-2 HLT validation suite in CMSSW**, enabling realistic performance evaluation of all key physics objects under HL-LHC conditions.
- **Demonstrated a working calibration-feedback prototype**, buffering data and reinjecting improved calibrations into the HLT scouting stream in real data-taking conditions.
- **Achieved two major L1 milestones**: a 40 MHz Phase-2 L1 Scouting prototype with

current technologies, and sustained operational deployment of ML-based anomaly detection in the Run-3 L1 Trigger.

Work Package 4 – Education, Outreach and Skills Development

- **Significantly expanded NGT visibility and community engagement**, through a high-impact communication strategy, a centralized project repository, and the successful 2nd NextGen Triggers Technical Workshop.
- **Delivered a broad, hands-on education programme**, including hackathons, workshops, tutorials, and exchange visits, directly supporting technical progress across WPs.
- **Advanced the CERN STEAM Academy to near-launch readiness**, defining its academic structure, partnerships, governance, and operational model for a 2026 start.

A photograph of two business professionals, a man and a woman, standing at a wooden table in a bright office. They are looking at and pointing to various business charts and documents spread across the table. The man is wearing a grey shirt and a black watch, and the woman is wearing a white shirt and jeans. The charts include a line graph, a pie chart, and a bar chart. There are also some sticky notes and paper clips on the table. A large red semi-circle is overlaid on the bottom left of the image, containing the text 'WPO: PROJECT MANAGEMENT AND COMMUNICATION'.

WPO:

PROJECT MANAGEMENT AND COMMUNICATION

Project Management and Communications

General Evaluation

The Project Management and Communications work package is responsible for the overall project coordination, the management of the contractual relations between CERN and Hillspire for the NextGen project, external partners and contributors, the internal CERN services, and CERN and Experiments management. The governance described in the Deliverable M1.0.1 and implemented at the beginning of 2024 has been applied also in Year 2 and has proven to be able to efficiently manage the project.

The Project Management Committee (PMC), chaired by the Project Coordinator and responsible for the project execution, and the Steering Board (SB), responsible for oversight and compliance with the CERN and experiments policies and objectives, have regularly met and managed the operations and strategy of the project leading to a successful conclusion of the second year.

In 2025 one of the main project management activities was the application of the plan defined at the end of 2024 to recover the recruitment and spending delays cumulated in 2024 for the reasons described in the Year 1 project report¹.

During 2025 the reprofiled recruitment plan was successfully applied leading to the completion of the expected workforce structure. As of December 2025, 31 new hires joined the project across the different activities. The recruitment schedule for 2026 has been further optimised to favour the hire of postdoc researchers earlier than previously planned and a faster turnover of expertise to impart an additional acceleration of the activities and improve the knowledge transfer across the community.

The reprofiled budget curve for 2025 and following year is being followed as planned. A break-even point between the original budget curve and the reprofiled curve is expected in 2027. The current curve foresees a 4% over-expenditure at the end of the project in 2028 as a mitigation for any additional unforeseen recruitment or retention issue and the volatility of the USD vs. CHF exchange rate.

Specific information on the communications activities is provided as part of the WP4.1 outreach report later in this document.

¹ <https://cernbox.cern.ch/s/NMuCwLRuBB2k1ip>

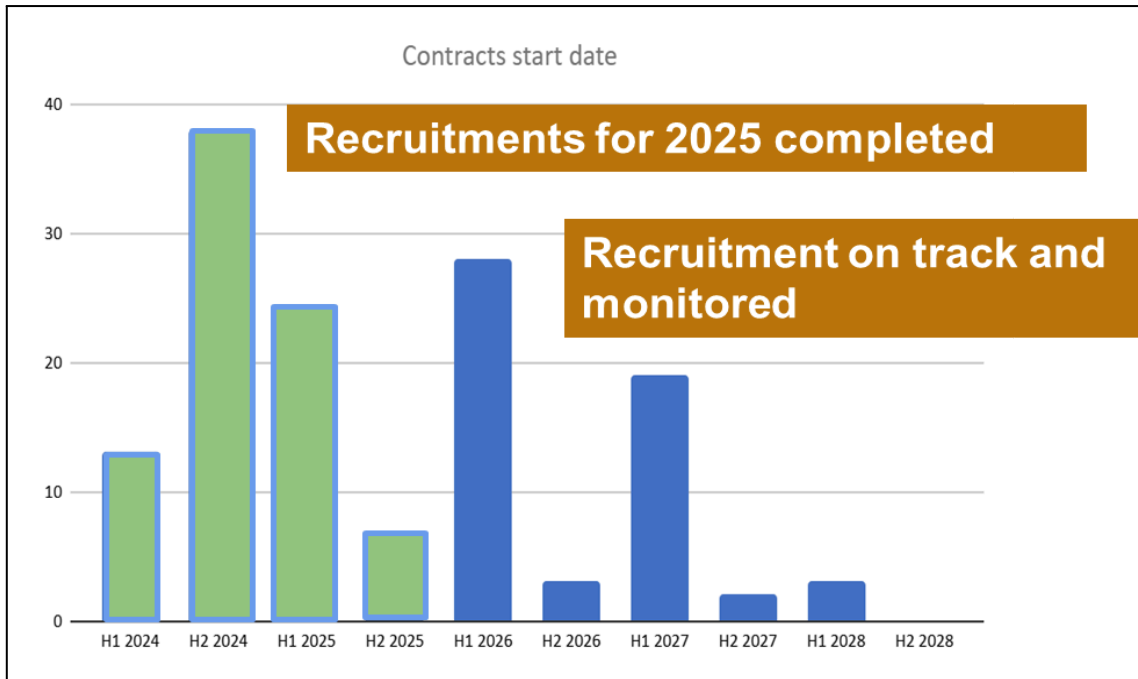


Fig. 1: End of Y1 (top) Vs. revised end of Y2 (bottom) contracts start dates distribution (head count, in **GREEN** the contracts already awarded in 2024 and 2025)

M2.0.1 Project management, risk management, activities and resources report

The main deliverable from the Project Management and Communications work package is the publication of the Year 2 Evaluation Report (this document) and the Expenditure Responsibility Report. In addition, the Project Management plan has been updated with new names and responsibilities. Due to changes in the CERN structure and governance as of January 2026, a new Project Coordinator and a new Deputy Project Coordinator have been nominated by the project Steering Board. The new Coordinators have started their mandate on January 1st, 2026. The new Project Coordinator is an existing experienced member of the NextGen Project Management Committee to guarantee a smooth transition without any disruption.

The Risk Management Plan and the Communications Plan developed in Q2 2024 have been applied and updated as necessary. In particular a new version of the NextGen Publications – Acknowledgment Guidelines has been developed to provide consistent ways of acknowledging the Fund and maximise the outreach and impact of the project activities. In 2025 thanks to NextGen the Eric & Wendy Schmidt Fund was also registered with the Open Funder Registry with funder ID 100032857 to

formalise the inclusion in scholarly metadata deposited with Crossref² and improve tracking of publications and impact.

The Periodic communication activity reports have been sent to Hillspire for information as foreseen by the Grant Agreement.

The following proofs of evidence are provided:

Deliverables

[Evaluation Report 2025 \(this document\)](#)

Expenditure Responsibility Report 2025

² <https://www.crossref.org/>



WP1:

INFRASTRUCTURE, ALGORITHMS AND THEORY

Work Package 1: Infrastructure, Algorithms and Theory

General Evaluation

Work Package 1 covers infrastructure, algorithms and theory, and plays a central role by providing the shared platform for R&D, common frameworks and computing resources that support all other Work Packages.

During the current reporting period, WP1 successfully established itself as the backbone of common R&D: monthly plenary meetings on machine-learning topics (coordinated by Tasks 1.2 and 1.3) attracted approximately 50 participants each time from across the project, encouraging cross-package collaborations. In addition, WP1 co-organized a series of technical workshops and events. These workshops saw participation beyond the core project teams, underscoring external interest and reinforcing the project's relevance to the broader high-energy physics and computing community.

The WP1 computing cluster is operational and used in production by project members for algorithm development, ML-based simulation and data-processing pipelines. The stable operation of that shared infrastructure continues to underpin the R&D and integration work required by the downstream work packages, notably towards real-time selection, high-performance computing and advanced trigger strategies envisioned for HL-LHC.

M2.1.1 Purchase of hardware and services and commissioning completed for on-premise and cloud resources

This work follows on the Y1 “Tender specification finalized and procurement launched for limited seeding resources” milestone, and includes:

- The order and purchase of the hardware previously tendered
- Validation, commissioning, integration of on-premises resources in CERN data centres
- Availability of the resources to users of the NGT cluster, along with commissioning of the resources in the CERN data centres, as well as the availability of external resources in public cloud providers

Availability of the hardware as well as the multiple interfaces available to end users are referenced directly in the [user documentation](#), along with the multiple flavours aiming to optimize overall GPU usage efficiency by partitioning resources to better suit individual workloads. Hardware is accessible interactively (via SSH, Notebooks,

VSCoDe, and others), in a batch mode of operation (with MPI support) and via an extensive MLOps platform for training, hyper-parameter tuning and serving workloads. Extensive monitoring features are available to end users and project managers, including GPU usage and availability.

Public cloud resources were tendered and offers made via blanket contracts established by CERN with different providers, with 3 core areas considered:

- R&D on optimization of hybrid deployments, allowing the bursting of workloads between on-premises and external accelerators
- Validation of new hardware generations, complementing the existing on-premises resources
- Support for online challenges and disseminating activities, in particular for users without access to CERN resources

Resources will be available in Q1 2026.

Deliverables

[Report](#) on purchase of hardware and services and commissioning completed for on-premise and cloud resources

M2.1.2 hls4ml software release 1 with open-access documentation

A new version of hls4ml (v1.2.0) was released in 2025 (<https://github.com/fastmachinelearning/hls4ml/tree/v1.2.0>), as documented in the accompanying paper, posted on the arXiv <https://arxiv.org/abs/2512.01463> and submitted for publication in a scientific journal.

With this release, **hls4ml establishes itself as a full compiler stack**, going beyond its original role as a Keras-to-HLS conversion tool. The main novelties are broader framework support, improved quantization correctness, multi-project compilation capabilities, and robust ASIC-oriented industry flows. While hls4ml sees contributions from outside the Next Generation Triggers project, most of the key features for this new release have been developed through NGT contributions. **Novel features include:**

- **Multi-ModelGraph:** models can now be split into multiple smaller subgraphs, which can be converted individually by the HLS library and then stitched together in a pipelined design. This feature is particularly important for large models, for which FPGA firmware generation time had previously become prohibitive.
- **Bit-exact, model-level precision propagation:** the new optimizer performs global precision propagation across the entire model using interval arithmetic,

ensuring bit-exact behavior for fully quantized networks. This removes the need for heuristic or user-defined accumulator precisions and prevents silent overflows, enabling numerically reliable and reproducible hardware implementations.

- **Full Keras 3 and HGQ 2 support:** the new version natively supports Keras 3 and interfaces directly with the HGQ library. This enables fine-grained, weight-specific quantization, including sub-layer and per-parameter bit-widths, leading to more effective model compression and improved deployment efficiency on hardware triggers.
- **Support for transformer architectures:** a new multi-head attention layer has been added to the set of supported operators, enabling the deployment of transformer-based models using the Vitis backend.
- **A new Catapult HLS backend:** this backend enables the use of hls4ml for ASIC design workflows. Functionality comparable to FPGA deployment is provided, including both latency-optimized and resource-optimized strategies for several supported architectures, as well as hierarchical and bottom-up synthesis flows.
- **Altera FPGA support:** a new oneAPI backend allows users to target Altera FPGAs, extending capabilities previously available mainly for AMD/Xilinx devices through the Vivado/Vitis backends.

In addition, this release integrates several supporting tools that were previously unavailable or insufficiently mature. These include support for symbolic regression models beyond standard neural networks, as well as surrogate models to estimate hardware resource consumption and latency prior to full HLS deployment.

Deliverables

[HLS4ML Software Release 1](#), [Open-Source Documentation](#)

M2.1.3 Define and document new-physics scenarios to evaluate trigger performance. Develop and deploy quantum circuit simulations for large systems, up to $O(100)$ qubits, using tensor networks and state-vector.

For WP1.4, we advanced and deployed quantum-circuit simulation capabilities targeting large systems up to $O(100)$ qubits by integrating tensor-network methods. These developments enabled the accurate reproduction of scattering processes, string-breaking phenomena, and flux-tube dynamics with controlled truncation errors, while also validating the feasibility of hybrid TN workflows for near-term,

hardware-relevant circuit depths. The results are documented in [five reports](#). The resulting simulation suite constitutes the concrete deliverable for the 2025 milestone and provides a unified framework for large-scale simulations aligned with the scientific goals of work package 1.4.

Work package 1.6 kicked off with the “BSM benchmarks for NGTs” [workshop](#) on July 4th, bringing together the theory and experimental communities. ATLAS and CMS presented the expected improvements arising from NGT and a speculative brainstorming session followed, to identify interesting BSM signatures which could benefit from those improvements. Several discussion meetings followed; at the September ‘25 [workshop](#) three main directions for BSM benchmarks were identified, for which work towards quantifying the potential has started:

- Flavor physics: on one hand considering observables related to rare decays of B-mesons, where improvements in the triggering of soft final state leptons can bring increased sensitivity; on the other hand considering low-mass resonance searches, such as a 10 GeV Z' boson arising from a model of flavour deconstruction, resulting in signatures with multiple soft b jets.

- Dark QCD-like sectors, with signatures of emerging or displaced jets. Most searches targeting these signatures have focused on triggering on H_T , but it would be interesting to explore if adding track information at the trigger level could help in analysis targeting lighter mediators.

- Long-lived particles, with signatures of disappearing or late-appearing tracks. Most searches for these particles focus on triggers in missing energy, and it would be interesting to explore if adding track information could add more signal events to the analysis. In particular it would also be interesting to understand the limiting factors of the searches when interpreting them in terms of more complex models.

Deliverables

[Technical Reports](#)

[Workshop on BSM Benchmarks for NextGen](#)

M2.1.4 Workshop at CERN for discussing the status and plans for all of the sub-projects in WP1.7

The research and development in Task 1.7 has proven to be central to several other tasks within the projects, with regular cross-attendance of meetings and combined works. Examples include struct-of-array work in this Task that is under consideration by ATLAS, CMS, and ALICE; memory layout for fast machine-learning inference as accomplished in 2025 by this Task and applied to CMS; scheduling improvements that are currently under study for ATLAS’s ATHENA framework.

The Task also invested in the fast inference library SOFIE. The optimizations and features implemented in 2025 bring SOFIE on par or in the lead of the state-of-the-art inference libraries, with often smaller memory footprint and faster inference times on CPU. Extension to GPU and multi-threading support are planned for 2026.

The scheduling research line progressed very well, with a new staff researcher who started in 2025 and a more application-focused research that - besides the more theoretical analysis of techniques for heterogeneous scheduling (for instance coroutines and fibres) - analyzed and investigated potential improvements from such research for the ATLAS and CMS software.

A team of students, together with research staff from the Task, conducted a first study of the Mojo language (very similar to last year's setup for Julia) which turned out to be very promising. The study will be extended in 2026 to cover multi-threading behavior and GPU kernels written in Mojo. The work on Julia has been concluded with a [report](#).

The accelerated common library GenVectorX saw an extended CUDA backend and new HIP and Alpaka backends, including benchmarks for all the backends. The clustering library CLUEsterng was considerably improved in 2025. It was included in the CMS stack with testing ongoing for several applications in different reconstruction domains, showing its generality and usefulness for all the experiments at CERN. Since 2025, the common heterogeneous data structure library contains an implementation of an association map in the target backends (CPU, CUDA, HIP and Alpaka).

The work on event generators has culminated in a complete implementation of leading-order generation in MadGraph. Late in 2025, a new investigation on floating point precision has been onboarded. Additionally, new developments aimed at improving memory management in GPU kernels and its access syntax are ongoing, and they will make use of the expertise acquired on struct-of-array within this Task. First results are expected in 2026 with direct applicability to the generator tasks included in this Task and much wider relevance to many of the other tasks of this project.

The research team has also interfaced with parties outside of Next Generation Triggers, for instance EDG for their implementation of the upcoming standard C++ reflection facility and in general the ISO C++ Committee JTC1/SC22/WG21; through interaction at conferences such as ACAT and EuCAIFCon; and through many workshops. The Task held a [dedicated workshop](#), presenting its progress to the wider community. With an attendance of more than 50 people, vivid and constructive discussions, and many follow-up leads, this event turned out to be a big success for the Task and will likely be repeated.

Deliverables

[Workshop on WP1.7 activities](#)
[Report on WP7 Activities](#)



WP2:

**ENHANCING THE
ATLAS TRIGGER
AND DATA
ACQUISITION**

Work Package 2: Enhancing the ATLAS Trigger and Data Acquisition

General Evaluation

In the second year of the project, all Work Package 2 activities are in full swing. The staff, post-docs and students hired on the project are well integrated in the CERN team and the ATLAS working groups. Foundational software infrastructure has been developed for deploying novel machine learning algorithms at the Level 0 trigger and ACTS based Inner Tracker and Muon Spectrometer software is taking shape. R&D on novel Machine Learning trigger approaches show promising results. A tool to study trigger rates for different selection strategies has been put in place and a prototype L0Global board is in production with a VP2802 FPGA that features an AI engine. The software infrastructure is taking shape to study the enhanced VP2802 capabilities for the Level 0 trigger. The NextGen Triggers proposal to change the ATLAS Phase-2 readout architecture has successfully passed the collaboration review in October, making it the first NextGen Triggers result that leads to a change of the baseline upgrade for the experiment.

The reports on the contractual milestones **M2.2.1** (Task 2.1), **M2.2.2** (Task 2.4) and **M2.2.3** (Task 2.5) are discussed in the subsequent sections. All 2025 contractual milestone reports and WP2 summary talks given at the 2nd NGT Technical Workshop are [collected here](#). A [list of Work Package 2 documents](#) is given that are made public either as ATLAS public notes, journal publications or EDMS documents. A list of presentations on Work Package 2 results at international conferences, workshops and at ATLAS workshops is given as well.

Results and achievements of the other Work Package 2 tasks in 2025 that are not subject of a contractual milestone are briefly summarised here:

Enhancing the L0 Muon Trigger (Task 2.2)

The NGT Task 2.2 activity is focused on enhancing the L0 Muon trigger, particularly on using precision data from the monitored drift tube chambers (MDTs) in the first level of the trigger, a new capability for the HL-LHC upgrade. The two goals of Task 2.2 are: to improve the robustness of the muon trigger in case of reduced performance or coverage of the dedicated trigger detectors, particularly the Resistive Plate Chambers (RPCs), and to investigate dedicated triggers targeting signatures of exotic new physics including long-lived particles, to enhance current sensitivity and enable new analyses.

Significant progress has been made in 2025 towards these goals in several fronts. Working together with Tasks 2.5 and 2.7, we have made progress developing the

simulation in Athena, including implementing the L0 Muon trigger algorithm for MDTs for the first time and producing large statistics samples to develop new trigger algorithms using both simulated muons and data from Run 3. These samples have been used to develop and study the performance of several algorithms for pattern recognition using ML, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Neural Networks (GNNs) and Transformers ([ATL-COM-DAQ-2025-125](#)). We have investigated the recovery of missing triggers from the RPCs using only MDT information, including algorithms to determine the collision time (Bunch Crossing Identification or BCID). Finally, we have developed a Neural Network (NN) to improve the muon momentum estimation in the trigger, that outperforms simple parametrizations by accounting for effects such as the inhomogeneous toroidal magnetic field. Regarding the development of triggers for exotic signatures, we have produced several benchmark simulated samples together with Tasks 2.5 and 2.7 to study a range of displacements and di-muon opening angles. These samples have been used to train algorithms to perform pattern recognition using CNNs, and to perform a regression on the displacement. We are also investigating algorithms to identify signatures of nearby muons from the decay of highly displaced long-lived particles.

The focus of this task has been shifting from algorithm development increasingly to implementation in firmware to use on FPGAs in the trigger. In collaboration with NGT Tasks 1.2 and 1.3, and using resources provided by Task 1.1, several of the ML algorithms developed have been implemented in firmware and the inference tested in several accelerator cards using HLS4ML. We have also made major progress in the validation of the firmware and the implementation of a verification pipeline to test the various modules of the Muon L0 trigger electronics and investigate several edge cases. Finally, we have benefitted tremendously from the student visitor program supported by Task 4, which has enabled participation from several students contributing to NGT projects from institutes that were previously not engaged in the L0 Muon project.

High Throughput Data Collection (Task 2.3)

The NGT Task 2.3 activity focuses on optimising the performance of the ATLAS Readout system, with particular emphasis on server platforms, CPUs, and network interfaces. The work is carried out by NGT personnel in close collaboration with the ATLAS TDAQ project.

In 2025 the effort concentrated on the development of a unified readout architecture, which consolidated into a single server the functions previously distributed across two separate computing nodes in the baseline design. This architecture offers significant benefits: reduced system complexity, lower risk of failures, simplified maintenance, and slightly reduced power consumption. Notably, the total number of required servers decreases from nearly 1,000 to approximately 300.

NGT personnel contributed directly to the unified architecture effort through hardware procurement, installation, configuration, software development, and detailed performance and thermal measurements. They also led the consolidation of all results into a dedicated report, which formed the basis of a formal architecture review. During this review, held in October 2025, the unified architecture was adopted as the new ATLAS baseline.

The performance studies carried out in this context highlighted several key lessons: processing must remain within CPU cache boundaries; performance does not scale linearly with the number of concurrent processes due to network resource contention; and “zero-copy” techniques are essential to meet Run-4 throughput requirements.

Additional work in 2025 included the development and comparison of memory-allocation drivers (CMEM vs. CMA), evaluation of new network-interface technologies, RDMA developments, and the first steps toward supporting ARM-based systems. Overall, NGT Task 2.3 delivered substantial advances in readout architecture, performance evaluation, and infrastructure readiness for ATLAS Run 4 and the HL-LHC.

Common Tracking Event Filter Infrastructure (Task 2.6)

Substantial progress has been made in 2025 regarding the common infrastructure development for a shared EF Tracking and offline software ecosystem: for the EF Tracking technology choice all pipeline, independently if they are purely executed on a CPU or with hardware accelerators (GPGPUs and FPGAs) have been interfaced to ACTS or directly use ACTS components. In all technology flavours certain setups exist that perform the later stages of track reconstruction (most prominently the track fitting, but for other setups also the track finding stage) within ACTS.

Immense progress was also achieved in developing and establishing the so-called Gen3 geometry model for ACTS, which will - when finalized - enable both, the correct description of the ATLAS Calorimeters and Muon Spectrometer, and a native translation of the geometry into the GPU friendly compile-time polymorphic geometry model *deTray*. During the joint WP2 and WP3 hackathon in October, a prototype of the CMS detector in Gen3 was demonstrated, and the experimental Gen2 geometry version was removed from the repository in Dec 2025. The upcoming 2026 work goals foresee the shift to Gen3 as the sole geometry model for ACTS, and the final phase-out of the Gen1 legacy code.

The GPU R&D project, *traccc*, has undergone a big leap towards maturity in 2025, and with the end of the year the target reciprocal throughput of 100 Hz could be demonstrated on the Open Data Detector, the application on the ATLAS ITk has been successfully implemented turning the *traccc* based pipeline into a realistic possibility for

the ATLAS EF technology choice decision 2026.

Enhanced Reconstruction for Higher Level Event Filtering (Task 2.7)

The NGT Task 2.7 activity focuses on the identification of physics signatures which are difficult to trigger on, building trigger-based solutions for such cases and estimating the trigger rate/bandwidth required for such solutions, with the ultimate aim of building a realistic HL-LHC trigger menu with enhanced capabilities.

During 2025 WP 2.7 has made significant progress in defining the physics cases that will be tackled, technical implementation of analysis tools, emulating the rates of triggers in the conditions of the HL-LHC and menu design for Trigger Level Analysis (TLA).

Near the start of the year an ATLAS-wide survey was performed to identify high priority physics cases as well as unusual signatures that would not be collected by the usual single/di-object triggers. The highest physics priority was found to be Di-Higgs boson production with the channel most difficult to trigger on being the case when both Higgs bosons decay to 2 b-quarks ($HH \rightarrow bbbb$), and several exotic signatures of new physics were identified that require further study in the latter stages of the project.

On the technical side, a trigger analysis framework has been developed for the quick implementation of new triggers and the evaluation of the rate that needs to be assigned as well as the signal acceptance that they would bring. This will enable the rapid turn-around of trigger suggestions in the coming stages of the project, and also includes native ONNX support to enable the use of ML in the proposed signatures.

The trigger rates of hadronic events, such as those recorded by the $HH \rightarrow bbbb$ triggers, are hard to simulate and are currently determined from data which is not possible in advance. Methods of emulating these rates in simulation have been developed and they have been benchmarked against Run-3 data proving their effectiveness and the degree to which they can be relied upon.

Signatures with multiple muons have high rates but significant physics of hadron decays can be achieved using trigger-level-objects in TLA, although this hasn't yet been exploited in ATLAS. Rate estimates and performance of this analysis chain has been evaluated showing significant promise.

Finally, looking further into the HL-LHC timetable it is possible that when the inner detector inner layers are replaced tracking hits might be able to be read out. Prospects for using these for identifying b-quarks at the lowest trigger level has also been

investigated as an example of exploring beyond the baseline ATLAS HL-LHC trigger program.

M2.2.1: First integration of ML in L0 Global trigger for commissioning and preparation of further improvements

The ATLAS Phase-II upgrade for the High-Luminosity Large Hadron Collider (HL-LHC) presents an unprecedented data processing challenge: selecting scientifically valuable events in real-time from a 40 MHz collision rate with an average of 200 simultaneous interactions (pileup). Work Package 2.1 (WP2.1) is positioned at the forefront of this challenge, with a clear mission to explore, develop, and deploy novel Machine Learning (ML) solutions for the Level-0 Global Trigger (L0Global). Our mandate is to enhance the trigger's physics reach by embedding sophisticated, ML-based, offline-like reconstruction capabilities directly into the hardware, operating within a strict 10-microsecond latency budget.

To achieve this, WP2.1 is pursuing a multi-faceted strategy focused on building a sustainable and powerful ML ecosystem for the Global Trigger. The core goals of this initiative are:

- Develop a common software framework to streamline the deployment of ML algorithms within the complex Global Trigger hardware environment.
- Evaluate novel ML model architectures that are explicitly optimized for the unique constraints of the trigger system, balancing physics performance with power consumption, latency, and hardware resource utilization.
- Explore new industry solutions, such as FPGAs with integrated AI-Engines, to push the boundaries of real-time processing and explore opportunities to enhance the trigger's capabilities beyond the baseline design.

In 2025, the WP2.1 team was significantly strengthened by the arrival of four new members, bringing fresh expertise and dedicated effort to our core tasks.

The central component of our 2025 contractual milestone is the development and implementation of an automated pipeline for ML model deployment. This testbed (ML2FW) is engineered to fundamentally reshape the paradigm of trigger algorithm development, shifting from bespoke, manual firmware design to a rapid, automated, and physics-driven exploration of the vast ML solution space. By automating the complex steps from a software model to firmware IP, this framework allows our team to rapidly explore a vast phase-space of potential ML solutions.

The testbed's primary objective is to automate the entire process from a user-provided ML model to a hardware-ready implementation, accounting for both physics performance and hardware constraints.

The pipeline executes through several distinct, automated stages:

1. Hyperparameter-Aware Training (HAT)
2. Quantisation-Aware Training (QAT)
3. Physics-Aware Selection
4. Hardware Conversion and Build:

The software framework for each stage of the ML2FW testbed is now complete, and we have demonstrated the process by which the components work together in a seamless, generalized workflow. This powerful testbed is the engine that will drive our diverse algorithm R&D portfolio, enabling the exploration of multiple advanced ML techniques detailed in the following section. All the software and tools for the testbed can be found in the following repository:

<https://gitlab.cern.ch/atlas-nextgen-wp21>

A diverse R&D portfolio is essential for tackling the wide range of physics challenges faced by the L0Global trigger. The automated ML2FW framework enables us to pursue multiple algorithm development tracks in parallel, each targeting a specific opportunity to enhance ATLAS's physics discovery potential. This section details the primary ML projects currently under development within the NGT WP2.1 purview, and which may ultimately serve as practical applications and validation cases for our core infrastructure:

- Jet-Aware Training for Self-Assembling Triggers
- Electron/Gamma (e/gamma) Classification
- Tau Identification (Seeding and Classification)
- Large Radius (Large-R) Jet Tagging
- Calorimeter-Based Vertexing
- Unsupervised Anomaly Detection
- Binary Neural Networks (BNNs) for FPGAs

While our primary development targets the baseline L0Global hardware, WP2.1 is also pursuing a forward-looking strategic initiative to evaluate technologies that could offer a significant performance boost. The GCM-AI project is designed to assess the viability of next-generation FPGAs featuring integrated AI co-processors (AI-Engines) for the demanding, low-latency environment of the ATLAS trigger. This initiative serves as a strategic exploration of the potential for upgrading the baseline FPGA architecture and hardware, thereby potentially reinforcing the processing capacity of the ATLAS hardware-based triggers system and ensuring that the experimental collaboration is positioned to exploit next-generation commercial technology in the HL-LHC era.

The project involves the study of a strategic modification to the Global Common Module (GCM) design in which the baseline AMD Versal Premium VP1802 FPGA (where all algorithmic processing for feature extraction and event selection is to take place) is replaced by the pin-compatible VP2802 variant, which includes a large array of AI-Engines (AIEs). These AIEs are fixed-latency vector processors optimized for ML workloads. Preliminary simulation studies performed by the WP2.1 team showed promising results for BDT and CNN implementations on the AIEs, suggesting it may be feasible to offload complex computations from the FPGA's programmable logic to the AIEs, thereby freeing up critical resources and potentially reducing overall latency.

A critical feasibility study was conducted to analyze the power consumption of the VP2802. Our analysis concluded that it is a viable option for the GCM, with an estimated power draw ranging from approximately 143W (without AIEs active) to 172W (with AIEs at maximum utilization), which is within the operational envelope of the system.

We are pleased to report that NGT will fund the production of GCM-AI, a GCM v4.0 prototype board hosting the advanced VP2802 FPGA. The FPGA has been successfully procured and delivered. We are now waiting for the opportunity to assemble the GCM-AI.

This hardware initiative is enabled by a crucial collaboration with Brookhaven National Laboratory (BNL). The BNL team, a key partner in the GCM development, will assist with the production and assembly of the GCM-AI board and provide invaluable expertise during the critical power-up and validation phases.

Deliverables

[Testbed Software](#) (with [technical report](#))

M2.2.2: Prototype GNN tracking algorithm based on ACTS

Work Package 2.4 aims to develop the most effective and performant solution for the ATLAS Event Filter (EF) tracking, employing both optimal classical numerical and machine learning (ML) techniques, and deploy it on the best fitting hardware architecture. It aims at enhancing both the physics and processing performance of track reconstruction, exploiting the novel tracking infrastructure via the use of ACTS, developed in Work Package 2.6 (Common Tracking Event Filter infrastructure), and investigating the feasibility of utilizing systems with co-processors like GPUs and FPGAs for parts of the track reconstruction. The personnel involved in WP 2.4 are fully integrated into the relevant ATLAS software and reconstruction activity areas, where they are developing and deploying complete pipeline solutions for EF Tracking. These solutions will be documented in dedicated ATLAS internal reports by the end of 2025,

ahead of a final decision on the architecture of the ATLAS EF tracking farm, planned for 2026.

Below is a summary of the main activities carried out:

- For the EF Tracking CPU pipeline, the full ACTS-based reconstruction chain was deployed in athena for the Event Filter. Significant speed-ups were achieved for both offline and online tracking, with extensive optimisation of the seeding and track-finding algorithms and their configurations. Detailed studies of the tracking performance of the ACTS-based EF solution were conducted, and substantial contributions were made to the results presented in the C100/C230 EF Tracking pipeline report. Final results are reported in the [ATLAS TDAQ EF Tracking CPU Technology Final Report](#).
- For the EF Tracking FPGA pipeline, FPGA kernels for data preparation within the F100 pipeline were developed, validated, and deployed. Physics and computational performance studies were carried out under realistic resource constraints. This work contributed significantly to the results documented in the F100 EF Tracking [pipeline report](#). Final results are reported in the [ATLAS TDAQ EF Tracking FPGA Technology Final Report](#).
- For the EF Tracking GNN pipeline, the ACTS-based GNN workflow was developed, validated, and deployed. Dedicated optimisations of its components were performed, including investigations of alternative inference libraries. R&D efforts addressed the reduced hit efficiency in the strip detector, exploring the use of GNN-based track candidates from the pixel detector and their combination with CKF. These activities formed a central part of the results presented in the G300 EF Tracking [pipeline document](#). Final results are reported in the [ATLAS TDAQ EF Tracking GPU Technology Final Report](#). A demonstrator of the ACTS-based GNN workflow has been developed.

Deliverables

[Software and Demo](#)

M2.2.3: Prototype ML and ACTS based muon reconstruction algorithm

The focus of work package 2.5 has been in two areas: to implement muon reconstruction in the common ACTS tracking framework and to use machine learning algorithms to accelerate the muon reconstruction process.

Muon reconstruction begins with pattern finding via a Hough Transform, followed by reconstructing straight-line segments in the individual chambers of the muon

spectrometer and concluding with the reconstruction of tracks from the reconstructed segments. Algorithms for all stages are implemented in ACTS and are then called from the ATLAS athena reconstruction code (see links below). The initial performance of the prototype algorithm chain is documented in [\[MDET-2025-555\]](#), with the highlights being:

- Efficiency of 98% for reconstructing muon segments within chambers (when the muon left at least four hits in the chamber) across all parts of the detector.
- Angular resolution of segments is found to be better than 1mrad across the full detector in the direction in which the muons have a curved trajectory. This high resolution is critical for achieving good resolution on the momentum of the muons.
- The efficiency to find seeds for track reconstruction is above 98% in all regions of the detector apart from those with reduced numbers of muon chambers.

The pattern finding and segment reconstruction code has also been migrated into the ATLAS EF framework to test the CPU performance [\[MDET-2025-555\]](#). The prerequisite of this was to update the configuration of the EF framework to use the new ACTS geometry model (see below for software links). The pattern finding and segment reconstruction is found to be 2.5 times faster than the equivalent reconstruction steps used in run-3. The reconstruction is also producing less segments that are not associated with a true muon and this results in a significant speed increase in the run-3 track building algorithm taking less than half the time when it is seeded with the segments from the new ACTS reconstruction.

The described performance marks the completion of the milestone for the prototype ACTS muon reconstruction. The prototype reconstruction chain will undergo further development and optimisation in the next year, including improving the tracking efficiency in the endcap region, full implementation of passive material into the ACTS tracking geometry and optimisation of the track fit.

The ATLAS muon spectrometer is characterised by having a large fraction of hits which do not originate from muons that come from the proton-proton collision. This large hit multiplicity is a significant driver in the CPU cost of the reconstruction algorithms and motivates employing machine learning methods to filter out the background hits. The first prototype algorithm for this task is the Muon Bucket Filter, which is available in the athena reconstruction code (see links below). The muon bucket filter processes “buckets”, which are groups of nearby hits in the muon spectrometer [\[ATL-DAQ-PUB-2025-004\]](#). It uses a Graph Neural Network (GNN) to evaluate if each bucket has 0, 1 or more than 1 muon in the bucket. Each bucket within an event is a node in the graph and edges between nodes are created if the nodes are spatially proximate. The performance of the GNN is tested in simulated samples. At the chosen working point, it is found to correctly identify more than 99% of buckets containing at least one true muon, while rejecting more than 96% of buckets which contain only

background hits. The impact of the GNN is tested by running the prototype ACTS reconstruction chain discussed above using only the buckets identified by the GNN as containing at least one muon. In these tests, the GNN runs on a GPU, while the rest of the reconstruction runs on a CPU. In simulated events with HL-LHC conditions, the set of algorithms including the GNN is found to run 15% faster than the set of algorithms without the GNN. This completes the milestone for prototype ML-based muon reconstruction.

Additional work beyond the milestones has been carried out to attempt full reconstruction of muons using a machine learning technique [[MDET-2025-04](#)]. The model runs in two stages, first filtering out background hits and then providing hit-to-track assignment and track parameter estimation. The performance of the filtering is strong, improving the hit purity from 0.6% to 67%. The track reconstruction efficiency is also high (98%), while the parameter estimates are currently worse than the standard tracking algorithms described above. This work provides a starting point for further developments of machine learning algorithms for muon reconstruction.

More details are given in the [M2.2.3 Milestone Report](#) and a report on the [software demo](#).

Deliverables

[Technical report](#)

[Software and Demo](#)



WP3:

RETHINKING THE CMS REAL-TIME DATA PROCESSING

Work Package 3: Rethinking the CMS Real Time Data Processing

General Evaluation

Work Package 3 focuses on enhancing the CMS Online Selection and Data Scouting workflows for HL-LHC by ambitious R&D for the L1 Trigger (L1T) and High Level Trigger (HLT). These efforts aim to remove the bottleneck of real-time event selection, extending the discovery and precision measurement capabilities of the CMS collaboration.

In 2025, the activities devoted to the HLT achieved two major milestones of the project's second year. The milestones **M2.3.1** and **M2.3.2** marked a major step forward in the R^3 and calibration-oriented developments of the NextGen project for CMS. These two deliverables addressed complementary aspects of the Phase-2 HLT challenge: ensuring that reconstruction performance can be measured with high fidelity under realistic conditions, and demonstrating that improved calibrations can be produced and fed back into the trigger system within operational time constraints. Milestone **M2.3.1** focused on the creation of a comprehensive validation suite, integrated directly into CMSSW, capable of evaluating the performance of the R^3 reconstruction across the full spectrum of key physics objects and representative signals. While this activity closed longstanding gaps—particularly in the validation of Phase-2 HLT objects and workflows—it was able to build on the solid foundations already present in CMSSW, extending and unifying existing tools where needed to deliver a coherent and robust validation framework. Milestone **M2.3.2** complemented this by delivering a functional small-scale prototype of the NextGen calibration-feedback system: a demonstrator capable of buffering data for O(8 hours), deriving improved calibrations, and applying them to the re-HLT reconstruction of scouting data. Despite operating on a reduced ($\approx 1\%$) input fraction, chosen to match the available hardware, the demonstrator successfully exercised the full architecture required for real-time calibration updates at scale. Together, the two milestones establish both the tools to evaluate reconstruction quality and the technical proof-of-concept for a future system capable of dynamically improving it, thereby advancing the NextGen vision and demonstrating readiness for full-scale Phase-2 integration.

In parallel, [Task 3.1.2](#) delivered significant advances in the CMS data-model infrastructure, extending the generic SoA framework with new composition mechanisms, AoS interoperability, and zero-copy data aggregation, thereby improving performance portability and enabling more efficient GPU-centric workflows within

CMSSW. [Task 3.2](#) advanced the distributed HLT execution model by consolidating an MPI-based client-server architecture, adding filter-aware data exchange, metadata-driven communication, and improved serialization strategies, and demonstrating a realistic small-scale distributed HLT prototype operating across CPUs, GPUs, and network boundaries. Finally, [Task 3.3](#) addressed RAW data volume constraints through systematic studies of lossy and lossless compression techniques, including the evolution from Raw Prime to Raw Second, extensive benchmarking of ROOT compression algorithms and formats, and exploratory GPU-accelerated compression, collectively establishing viable strategies to reduce HLT output size while preserving physics performance.

On the other hand, the NGT activities devoted to the L1T also achieved two major milestones of the project's second year. These milestones advance the long-term vision of enabling real-time physics at 40 MHz and integrating robust machine-learning-based algorithms into the L1T operational workflow. Milestone **M2.3.3** marked a crucial step in the development of the Phase-2 L1 Scouting system. The team successfully demonstrated the first working prototype capable of performing data acquisition and real-time physics analysis at 40 MHz using current-generation technologies such as Virtex Ultrascale+ FPGAs with High-Bandwidth Memory, 100 Gb/s networking, and commercial CPU/GPU processing, as well as delivering the first conceptual design of the next-generation ATCA data acquisition board based on Versal HBM devices and 400 Gb/s networking. These results confirm the technical feasibility and scalability of the full Phase-2 architecture foreseen for HL-LHC. Milestone **M2.3.4** instead concerns the real-time use of machine learning in the L1 Trigger during LHC Run 3. Throughout 2025, the team gained unprecedented operational experience with continuous training, calibration, monitoring, and deployment of ML models in a real collider experiment environment, centered around the anomaly-detection algorithm AXOL1TL. In addition, substantial developments were performed toward a robust MLOps infrastructure suited for a trigger system with tens of ML models deployed in parallel. The detailed reports below demonstrate the successful completions of **M2.3.3** and **M2.3.4**.

In parallel, the **Task 3.6** team continued working on the design and firmware synthesis of new low-latency ML-based algorithms with the goal of further improving the physics performance of the Phase-2 L1T event reconstruction. A major development is the integration into the Correlator Layer-1 firmware of the multi-class BDT designed in the previous year for the identification of High Granularity Calorimeter clusters. Moreover, electron reconstruction was further enhanced through a dedicated BDT-based transverse momentum regression, yielding a substantial improvement in transverse momentum resolution. These improvements are particularly relevant for low-mass resonance searches using the Level-1 Scouting system, where accurate electron kinematics are critical. As a major step forward, building on work initiated in the

previous reporting period, a new multi-class jet tagging neural network was also delivered. During 2025, the jet tagger was fully integrated into the Correlator Layer-2 firmware, including constituent sorting, preprocessing, and neural network inference. This implementation demonstrates the capability to enhance trigger acceptance for challenging signatures such as di-Higgs boson production in the four b-jets final state. All supporting materials detailing and demonstrating the outcomes achieved are provided in the dedicated [2025 Task 3.6 Report](#), which includes links to all presentations and documents produced throughout the year.

M2.3.1 A validation suite that accurately measures the performance of the R^3 reconstruction for key physics objects and representative physics signals under realistic data taking conditions is developed and integrated in CMSSW

The milestone for 2025 has been fully achieved. Building on substantial validation infrastructure already present in CMSSW through longstanding CMS developments, as well as more recent advances—especially in the area of the High Granularity Calorimeter (HGCal) and its reconstruction framework, TICL—the work carried out in 2025 significantly extended, unified, and integrated these tools precisely in the areas where validation capabilities were previously missing or incomplete, most notably for the Phase-2 High-Level Trigger (HLT) and the full suite of physics objects reconstructed at that stage. In parallel to the development of the validation tools, major progress was achieved across the project: the introduction of the NGT Scouting stream, serving as a dedicated playground to test more aggressive and innovative reconstruction strategies at the HLT; its storage in a lightweight, analysis-oriented NanoAOD-style format, enabling for the first time a realistic estimate of the expected output size in kb/event; and substantial reconstruction improvements, including the extension of pixel tracking to the first three macro-pixel layers of the Outer Tracker, yielding sizable gains in efficiency and fake-rate reduction, as well as a streamlined muon reconstruction chain that significantly improves organization and timing performance. As documented in the [2025 Task 3.1.1 Report](#), a comprehensive and unified validation framework has been designed, implemented, and deployed across all major Phase-2 HLT-reconstructed objects. The suite provides consistent truth definitions, robust performance metrics, and end-to-end validation workflows under realistic HL-LHC (PU = 200) conditions, and it is fully integrated into CMSSW, supporting both Phase-2 HLT development and the NGT Scouting stream. Its successful deployment has already enabled the assessment and validation of major reconstruction upgrades — prominently the Patatrack pixel-tracking extensions and the streamlined muon reconstruction — demonstrating that the required validation capabilities are not only delivered but actively amplifying reconstruction progress. Despite the scale of the achievement, the validation suite is expected to continue evolving in the coming years, expanding and adapting in parallel with future

reconstruction developments to ensure comprehensive coverage throughout the project's lifetime. All supporting materials detailing and demonstrating the outcomes achieved are provided in the dedicated [2025 Task 3.1.1 Report](#), which includes links to all presentations and documents produced throughout the year.

Deliverables

[Validation Suite Software](#)

M2.3.2 Creation of a small-scale prototype that buffers 30% of the RAW data for about 8 hours, derives improved calibrations, and uses them for the online reconstruction of the “HLT scouting” data stream

The milestone for 2025 has been fully achieved. A complete small-scale NGT Demonstrator has been designed, deployed, and commissioned, implementing the full functionality required by the milestone: RAW data buffering for O(8 hours), derivation of improved calibrations, and their use in the online reconstruction of the HLT scouting stream. For the purposes of validating the architecture and calibration-feedback loop within the constraints of the available Run-3 hardware, the system was configured to buffer and reprocess an O(1%) prescaled subset of the Level-1/HLT input—about 1 kHz out of a typical 110 kHz L1T output—while preserving the very same data flow, software chain, and timing constraints that would be required at higher fractions of the input rate. As described in the [2025 Task 3.4 Report](#), the demonstrator uses two 30 TB NVMe SSDs on a dedicated Storage Manager node to provide an 8-hour circular buffer, an automatic conversion and file-broker chain to feed the re-HLT, and a dedicated NGT HLT menu and Global Tag to consume updated conditions. The calibration leg was successfully operated end-to-end for ECAL laser transparency corrections, ECAL pedestals, and silicon strip tracker bad-component maps, with payloads derived within the buffering window, uploaded to the CMS Conditions Database, and subsequently picked up by the delayed re-HLT reconstruction of the scouting stream. The full chain—including data buffering, calibration production, conditions upload, and re-HLT processing with improved calibrations—was demonstrated in collisions in late 2025, completing the validation of the demonstrator infrastructure. A dedicated DQM framework for scouting has been developed and integrated, and although its physics-performance studies are ongoing, it is now ready for extensive use in production data-taking next year. In this way, a functioning small-scale prototype embodying all the key ingredients of the 30%-buffer concept has been realised, thereby fulfilling the 2025 contractual milestone and providing a validated blueprint for scaling up to higher input fractions in future years. All supporting materials detailing and demonstrating the outcomes achieved are provided in the dedicated [2025 Task 3.4 Report](#), which includes links to all presentations and

documents produced throughout the year.

Deliverables

Prototype Demo

M2.3.3 First prototype Phase-2 L1 Scouting system demonstrating data acquisition and real-time physics analysis in simple final states with present technologies (i.e. Virtex Ultrascale+ HBM, 100 GbE networking, current CPU/GPUs), and first conceptual design of next generation ATCA data acquisition board for L1 Scouting (Versal HBM, 400 GbE). Results documented as CMS public notes and conference talks/proceedings

In 2025 the NGT Task 3.5 team completed the first fully functional prototype of the Phase-2 L1 Scouting system, demonstrating for the first time sustained data acquisition and real-time processing at 40 MHz using present-day FPGA, networking and computing technologies. Building on the architectural groundwork from 2024, the 2025 demonstrator integrated a DAQ800 board equipped with two VU35P Ultrascale+ FPGAs with HBM and capable of handling up to 48 L1T input links at 25 Gb/s, synchronised via the TCDS2 protocol and streamed through seven 100 Gb/s TCP/IP outputs. The online processing cluster received and aggregated up to 140 Gb/s of particle flow candidates from the primary vertex, electromagnetic objects and muon candidates, running the suite of simple physics analyses developed in 2024 – including rare W and Higgs decay channels and low-mass dielectron resonances – as realistic benchmarks to validate the system’s end-to-end performance under Phase-2-like conditions. Compared to the 2024 setup, the 2025 demonstrator achieved significantly improved stability and scalability, operating simple analyses at the full 40 MHz bunch-crossing rate and deploying the entire DAQ and CMSSW workflow under Kubernetes for robust orchestration and monitoring. In parallel, the team delivered the first conceptual design of the next-generation data-acquisition hardware for L1 Scouting. The team evaluated 400 Gb/s networking technologies, demonstrated >390 Gb/s data transfer from FPGA to servers using RoCEv2, and defined a modular rack-mount architecture based on Versal Premium devices with LPDDR4 memory and QSFP-DD optical links. This design represents the evolution path beyond the current DAQ800 platform and satisfies the second component of the milestone. All supporting materials detailing and demonstrating the outcomes achieved are provided in the dedicated [2025 Task 3.5 Report](#), which includes links to all presentations and documents produced throughout the year. Together, these achievements fully realise Milestone M2.3.3, establishing the technological feasibility, performance scalability and hardware roadmap required for the Phase-2 L1 Scouting programme.

Deliverables

Prototype Demo

M2.3.4 Report on operational experience and achieved physics performance for the continuous training and deployment of ML algorithms in the L1 Trigger on Run 3

During 2025, the NGT Tasks 3.7 and 3.6 teams jointly achieved the objectives of Milestone M2.3.4 by consolidating a full operational cycle of training, validating, deploying, monitoring, and recalibrating machine-learning algorithms for anomaly detection running in the Level-1 Trigger during Run 3. Building on the deployment of the AXOL1TL anomaly-detection trigger in 2024, the team operated the upgraded V5 model throughout the 2025 data-taking period, gaining real-world experience in managing ML algorithms under the rapidly evolving detector conditions characteristic of LHC running. The studies performed on the anomalous events collected during the year confirmed the stability and physics relevance of the selections, showing that AXOL1TL efficiently targets regions of phase space characterised by low hadronic activity, low missing transverse momentum, and high object multiplicity. These datasets also enabled the start of model-independent bump-hunt searches based on L1 anomaly-triggered events, providing an initial pathway toward assessing the physics reach of this novel triggering strategy. To sustain stable operations over long data-taking periods, the Task 3.7 team developed and exercised workflows for both full retraining and fast recalibration of the anomaly-score distribution, ensuring consistent behaviour of the deployed model across changes in beam conditions, detector performance, and data-taking configurations. In parallel, Task 3.6 delivered the technological foundations of a scalable MLOps framework tailored to CMS triggering needs, including versioning and validation tools, firmware–software compatibility checks, automated deployment procedures, and continuous performance monitoring. These tools are designed to generalise the operational workflows tested with AXOL1TL to the much larger set of ML models foreseen in the Phase-2 L1 Trigger. The combination of real-data stress tests from Task 3.7 and the scalable infrastructure developed in Task 3.6 ensures that CMS is now building both the operational experience and the technological foundations required to enable continuous training and deployment of ML algorithms in real-time trigger conditions. All supporting materials documenting these achievements are provided in the dedicated [2025 Task 3.6](#) and [Task 3.7 Reports](#), which include links to all presentations and written documentation produced throughout the year. Together, these developments fully realise Milestone M2.3.4 and establish a solid foundation for ML-based triggering in

Run 3 and beyond.

Deliverables

[Technical report 1](#), [Technical report 2](#)



WP4:

**EDUCATION
PROGRAMMES
AND OUTREACH**

Work Package 4: Education Programmes and Outreach

General Evaluation

The work package 4 has clearly focused its 2025 strategy on increasing the visibility, accessibility, and recognition of its two tasks (Communication/Outreach and Education) across CERN, the high-energy physics community, and the general public. These efforts demonstrated the scientific and technical impact of NGT for future detector designs and showcased the support provided by the Eric & Wendy Schmidt Fund for Strategic Innovation.

Several parallel initiatives have been taken as detailed in the reports. Among them, we have the website, which is now the main hub for news, resources, and constantly updated multimedia (videos in particular). Over the year, the website published articles, interviews, insights from events and the NextGen Repository, a centralized tool for collecting all project outputs-papers, presentations, and conference talks - which gives a glimpse of the project's broad engagement and visibility across the scientific community.

The workpackage maximized its outreach by publishing content on multiple social media through CERN's high-visibility communication channels of CERN Computing, ATLAS, and CMS, which together have a multi million audience of followers. Communication efforts also included the NextGen Technical Workshop with three-day of highly technical presentations that summarizes the 2025 milestones, achievements, and coordinate future developments

Also the education activities have been proactive through a variety of focused education and community events and schools which have been proposed to all NextGen collaborators. This was coupled with specialized workshops and hackathons focusing on technical topics with tenths of cross-team contributors. for exchange and rapid prototyping. In addition to these organized events, specialized tutorials were delivered, and the NGT Learning Hub continued to publish reusable educational materials, sustaining continuous skills development and knowledge sharing within the NGT community and the wider LHC experiments.

Significant advancement was made in building the CERN STEAM Academy which should start during the summer of 2026. The Academy was defined as a multiple week summer university hosted at CERN, around three themes all relevant to NextGen technologies and skills.

M2.4.1 2nd NextGen Triggers Project Workshop. Report on exchange and outreach activities

Outreach activities

In 2025, the outreach efforts for the Next Generation Triggers project continued to expand, building on the communication foundations established in the previous year. The [NGT website](#) remained the central hub for all public-facing information, with more than [30 new articles](#) published throughout the year, providing regular updates on project activities, events, and technical developments. In parallel, over 20 videos were produced and shared across the website and CERN communication channels, significantly increasing the project's online visibility and engagement.

In addition to this, throughout the year, NextGen Triggers has organized and collaborated on several events, the main one being the three-day [NGT 2nd Technical Workshop](#) held at the Globe of Science and Innovation. This year also saw a significant increase in the participation of NGT members in conferences, workshops, and external events, where they presented their work through talks, posters, papers, and demonstrations. To support this growing output, a new [NGT Repository](#) was created, together with a dedicated project community, providing a centralized space for sharing and archiving all contributions and materials from across the collaboration.

Overall, the project has continued to broaden its outreach and strengthen its communication channels, ensuring that the developments and impact of NextGen Triggers remain accessible to a wide audience.



The CERN NextGen Triggers 2nd Technical Workshop attendants, November 2025.

Exchange programme

Work Package 1 benefited from the exchange program by organizing international workshops together with key experts. Task 1.5 organized such a technical workshop for expert exchange on the [suppression of negative weights](#). Task 1.7 organized several workshops on topics such as [reduced floating point precision](#) with 150 registered participants. As a significant side effect, the exchange program improved project hires, for instance by WP1 being able to attract a PhD student of one of the invited experts.

Work Package 2 has used Exchange Programme funds in 2025, to invite students from institutes not yet involved in the program but interested in NGT activities. We consider this approach the most favourable to attract new institutes into our activities. [Two master students from Uppsala University](#) have contributed to studies of Muon signatures from the decay of long-lived particles. During their two-week stay they have collaborated with WP2.2 to convert the simulated models into FPGA firmware. The outcome of this work is part of WP2 deliverables and master theses of these two students. The second Exchange support was used to invite a PHD student from Genoa, to work on filtering hits originating from pile-up interaction, prior to any track reconstruction. This approach is interesting for HL-LHC tracking, to limit the number of hits to be considered in track reconstruction, an active R&D subject for WP2.3.

Work Package 3 has been supported by the Exchange Programme funds in several activities. From December 1st to December 5th 2025, members of the Next Generation Trigger (NGT) project, together with external collaborators, participated in the LUMI Hackathon held in Lugano.

The focus of the activity was the optimization of the CMS Software for AMD GPUs, targeting in particular the MI250X available on the LUMI HPC system and the MI300X and Radeon Pro W7900 deployed on the NGT GPU farm at CERN. Through close and continuous collaboration with experts from LUMI and AMD, the team identified several effective optimizations: tuning environment variables to match specific hardware and software configurations, and adjusting the number of concurrent GPU queues used by the software.

In line with its mission to foster collaboration and knowledge exchange, WP4 sponsored the participation of NGT members as well as Abdularham Al Marzou (University of Bahrain), a former collaborator on projects involving NVIDIA and AMD GPUs, who joined the activities at CERN and at the Hackathon.

Deliverables

[2nd NextGen Technical Workshop](#)
[Outreach and Exchange Report](#)

M2.4.2 Skills gaps analysis done. Report on first year of the STEAM Programme activities

In 2025, WP4.2 advanced the education mission of the Next Generation Triggers (NGT) project to strengthen the community's capabilities in advanced software, real-time computing, and ML for next-generation trigger and DAQ systems. The work combined delivery of coordinated training activities and community educational events with structured programme-building to support sustainability and scale. In parallel, the CERN STEAM Academy progressed from concept to an advanced stage of readiness for launch.

The Skills Gap Analysis was already completed in 2024 and reported as part of the M1.4.2 report linked below.

Educational activities and community engagement

Across 2025, a set of education and community events was organised and supported to strengthen technical capabilities and promote exchange across NGT and the experiments. Selected activities included the [1st NGT Hackathon \(7–11 April 2025\)](#) focused on WP1 topics, the [ATLAS & CMS MLOps Workshop \(2 June 2025\)](#) addressing shared infrastructure challenges for ML deployment, the [NGT-Openlab “Optimising Floating Point Precision” Workshop \(1–2 July 2025\)](#), and the [TH-NGT Hackathon \(23–26 September 2025\)](#) supporting onboarding to the NGT Cluster. This was followed by the [2nd NGT Hackathon \(20–24 October 2025\)](#), bringing together 40+ contributors from NGT and the LHC experiments for rapid prototyping and cross-team exchange. Additional specialised tutorials were also delivered, and reusable educational outputs continued to be published via the NGT Learning Hub. Engagement with external schools was pursued to complement internal training.

CERN STEAM Academy: major progress toward launch

A central achievement in 2025 was advancing the CERN STEAM Academy from concept to an advanced stage of readiness for launch. The 2026 edition of the Academy was defined as a 10-week, on-site cohort programme organised around three interleaved themes - Modern Software Technologies, Edge Computing, and Data Science & Machine Learning, with a teaching model combining lectures delivered by partner-university faculty, tutor supported hands-on labs, and a weekly seminar series featuring invited high-profile speakers. The academic programme was developed with the support of the Programme Committee and experts acting as theme coordinators, ensuring alignment with NGT priorities and a coherent learning path across themes.

During 2025, collaboration with partner universities was created (ETH Zurich, UCLM, University of Warwick, University of Lyon, and more), the administrative and privacy

baseline required for future applications was established, and core operational planning was confirmed, including the hosting model and on-site arrangements.

An open-access policy was confirmed, with publicly available materials and capacity-limited access for the CERN community to teaching sessions. Overall, 2025 delivered substantial programme-building work to enable the next steps toward opening applications and finalising the detailed timetable and teaching weeks.

Overall, 2025 delivered impactful education activities for the NGT community and substantial progress toward launching the CERN STEAM Academy. In 2026, the focus will shift from preparation to implementation and delivery of the Academy. In parallel, education and community events will continue to sustain skills development across NGT, while expanding the publication and reuse of training outputs through the NGT Learning Hub.

Deliverables

[Skills Gap Analysis](#)

[Report on STEAM Academy Programme Activities](#)

Conclusions and Further Recommendations

The NextGen Triggers project has successfully demonstrated, in 2025, that the ambitious vision of transforming real-time data processing for the HL-LHC through advanced computing, machine learning, and novel architectures is both **technically feasible and scientifically impactful**. Across all Work Packages, the project moved decisively from exploratory R&D toward **validated prototypes, production-grade software, and architecture decisions** that are already influencing the baseline designs of the ATLAS and CMS experiments. The integration of ML into low-latency trigger systems, the maturation of common software frameworks, and the establishment of scalable infrastructure and validation tools collectively mark a major step change in how real-time event selection can be conceived and delivered.

A key conclusion is that **end-to-end integration matters as much as algorithmic innovation**. The strongest results emerged where infrastructure, software frameworks, firmware, and physics use cases were developed together, enabling realistic performance evaluation under HL-LHC conditions. The successful adoption of NGT-driven architectural choices by the experiments confirms the value of this tightly coupled, experiment-embedded approach. Equally important, the project has shown that **ML can be operated reliably in trigger environments**, provided that robust MLOps practices, monitoring, and calibration workflows are treated as first-class design requirements rather than add-ons.

Several lessons learned stand out. First, **shared platforms and common libraries are powerful accelerators**: investments in common clusters, ACTS-based tracking, hls4ml, and validation suites paid off by reducing duplication and enabling rapid cross-fertilisation between ATLAS and CMS. Second, **early and continuous engagement with experiment governance** is essential to translate R&D into baseline impact; milestones that aligned with experiment decision points proved especially effective. Third, the project confirmed that **skills and community building are critical enablers of technical success**: hackathons, exchange programmes, and targeted training directly contributed to faster prototyping, better integration, and wider adoption of results.

Looking ahead, several recommendations and next steps emerge clearly. From a technical perspective, the priority should be to **consolidate and scale**: mature the most successful prototypes into production-ready components, extend performance and stress testing to full HL-LHC conditions, and prepare for technology-choice decisions foreseen in 2026. ML deployment pipelines, calibration feedback loops, and heterogeneous tracking solutions should be hardened for long-term operation, with particular attention to maintainability, reproducibility, and power efficiency. Continued exploration of emerging

hardware technologies should proceed in parallel, but with clear criteria for integration into experiment roadmaps.

At the programme level, the project should further strengthen its role as a **cross-experiment integration layer**, maintaining alignment between ATLAS and CMS while preserving flexibility for experiment-specific optimisation. Sustaining and expanding the education and outreach efforts, including the launch of the CERN STEAM Academy, will be essential to address long-term skills needs and ensure continuity beyond the project's lifetime. Finally, the demonstrated success of NextGen Triggers argues for embedding its methodologies, namely common infrastructures, ML-aware operations, and iterative prototyping, more deeply into the standard model of trigger and DAQ development for future large-scale scientific facilities.

In conclusion, NextGen Triggers has moved from vision to impact. The foundations laid in 2025 provide a solid and credible path toward fully exploiting the HL-LHC physics potential, while offering a transferable model for real-time data processing in data-intensive science more broadly.

Appendix 1

Milestones for the Coming Year 2026

A detailed look at the key milestones expected to be reached in the next phase of the project.

Year	Code	Milestones	Type
3	M3.0.1	Project management, risk management, activities and resources report	Report
3	M3.1.1	Demonstrator of project use cases deployed on the chosen MLOps platform. Demonstrator of hybrid infrastructure usage (on-premises and public cloud).	Report, Demo
3	M3.1.2	NNLO software release 1 with open-access documentation	Software
3	M3.1.3	Document first examples of LQFT code-performance improvement, e.g. w.r.t. signal-to-noise ratio in hadronic matrix elements, and first examples of event generator acceleration, addressing leading-order production processes	Report
3	M3.1.4	Produce a report on the status of all of the sub-projects from WP1.7.	Report
3	M3.2.1	Integration strategy for enhanced ACTS in common tracking software infrastructure	Report, Software
3	M3.2.2	Proof of concept and decision on technology in enhanced high throughput data collection	Report, Software
3	M3.2.3	Readiness of ML overall tracking approach and optimisation of system design	Report, Software
3	M3.3.1	A prototype of the HLT R3 reconstruction is integrated in CMSSW, and its performance is compared with that of the baseline online reconstruction	Demo
3	M3.3.2	A comparative analysis of the different data reduction approaches and their impact on the accuracy of the physics reconstruction is published	Report
3	M3.3.3	Completed portfolio of physics analysis studies for Phase-2 L1 Scouting in all final states and associated requirements for the scouting system, documented in CMS public notes, PhD theses and conference talks/proceedings	Report
3	M3.3.4	Completed end-to-end Run 3 anomaly detection analysis, physics results documented in CMS physics journal publication / conference talks / PhD theses	Report
3	M3.4.1	3rd NextGen Triggers Project Workshop. Report on exchange and outreach activities	Event, Report

3	M3.4.2	STEAM Programme opened outside NextGen and integrated in CERN wider education activities	Report
---	--------	--	--------