



**NexTGen**  
Next Generation Triggers

# EVALUATION REPORT

---

◆◆◆ 2024



## Contact information:

CERN  
Espl. des Particules 1/1211, 23 Genève  
Next Generation Triggers:  
<https://nextgentrigger.web.cern.ch/>  
Project Coordinator:  
Alberto Di Meglio  
Email: [alberto.di.meglio@cern.ch](mailto:alberto.di.meglio@cern.ch)

# Table of Contents

<b>Executive Summary.....</b>	<b>1</b>
<b>Introduction.....</b>	<b>2</b>
<b>Evaluation Objective(s).....</b>	<b>2</b>
<b>Evaluation Methodology.....</b>	<b>2</b>
<b>NextGen Annual Milestones Status Sheet.....</b>	<b>3</b>
Add cover image for Management work package.....	5
Project Management and Communications.....	6
General Evaluation.....	6
M1.0.1 Project management, risk management, activities and resources report.....	6
Work Package 1: Infrastructure, Algorithms and Theory.....	5
General Evaluation.....	5
M1.1.1 Tender specification finalized and procurement launched for limited seeding resources.....	5
M1.1.2 MLonFPGA community workshop, to identify needs from ATLAS and CMS on WP 2 and 3 as inputs to WP 1.2 and 1.3.....	5
M1.1.3 Workshop to identify and prioritize event generators' components suitable for acceleration, and develop LQFT benchmarking software tailored to hardware infrastructure.....	5
M1.1.4 Report on the preparatory work done in WP1.7 and the concrete work planned for year 2 of all of the sub-projects.....	6
Conclusion.....	7
<b>WP2: ENHANCING THE ATLAS TRIGGER AND DATA ACQUISITION.....</b>	<b>8</b>
Work Package 2: Enhancing the ATLAS Trigger and Data Acquisition.....	7
General Evaluation.....	7
M1.2.1 ML development toolkit for AI algorithms on FPGA at fixed latency for the ATLAS global trigger.....	7
M1.2.2 Initial algorithm for L0 Muon MDT with RPC and Tile seeding.....	7
M1.2.3 Review of identified physics scenarios for enhanced trigger.....	7
Conclusion.....	7
<b>WP3: RETHINKING THE CMS.....</b>	<b>8</b>
Work Package 3: Rethinking the CMS Real Time Data Processing.....	9
General Evaluation.....	9
M1.3.1 Report on the online reconstruction performance, addressing identified bottlenecks, proposing specific improvement solutions, and outlining necessary features for the generic CMS Structure of Arrays (SoA).....	9
M1.3.2 Report on the impact of RAW data compression and of their replacement with low-level reconstructed quantities (RAW').....	9
M1.3.3 Completion of first studies on physics analysis prospects for Phase-2 L1 data scouting in simple final states, evaluating physics reach and requirements on the data scouting architecture (bandwidth, processing power).....	9
M1.3.4 Implementation and integration of AI-based algorithms for the Run 3 L1 Global Trigger (including unsupervised algorithms for anomaly detection) and Phase-2 Correlator Trigger, including definition of operational practices for continuous training and deployment of models during Run 3....	9
Conclusion.....	10
<b>WP4: EDUCATION PROGRAMMES AND OUTREACH.....</b>	<b>11</b>
Work Package 4: Education Programmes and Outreach.....	11

<i>General Evaluation.....</i>	<i>11</i>
<i>M1.4.1 1st NextGen Triggers Project Workshop. Report on exchange and outreach activities.....</i>	<i>11</i>
<i>M1.4.2 CERN STEAM Programme governance and outreach channels in place.....</i>	<i>11</i>
<i>Conclusion.....</i>	<i>11</i>
<b>Conclusions and Recommendations.....</b>	<b>12</b>
Conclusions.....	12
Lessons Learned.....	12
Good Practices.....	12
Recommendations.....	12
<b>Appendix 1.....</b>	<b>13</b>
Milestones for the Coming Year 2025.....	14



## Executive Summary

The Next Generation Triggers project (NextGen in short) is a five-year collaboration across ATLAS and CMS (with contributions from LHCb and ALICE) and the Experimental Physics, Theoretical Physics, and Information Technology Departments of CERN to research and develop new ideas and technologies for the experiment trigger systems for HL-LHC and beyond. The project is funded by the Eric and Wendy Schmidt Fund for Strategic Innovation. After more than a year of preparation in 2022-2023, the project started in January 2024. It involves the effort of more than 100 researchers and engineers over its five-year duration to work on four interacting areas: (1) online data processing, modern computing architectures, novel algorithmic concepts, machine learning and the direct interplay of experimental approaches and theory simulation; (2) enhancing the ATLAS trigger and data acquisition to focus on improved and accelerated filtering and exotic signature detection; (3) rethinking the CMS real-time data processing to design a novel AI-powered real-time processing workflow to analyze every single collision produced in the LHC; and (4) designing novel education and training programmes to support the experiment research plans.

This document is the Year 1 Evaluation Report to the Fund and highlights the achievements of the project during its first 12 months of execution and the status of the agreed milestones and deliverables.

During 2024 the project has defined and put in place its governance and structure and produced more detailed plans for each of the activities. The main challenge this year has been to recruit experts in the numbers and level of competence required by the ambitious programme. Despite some delays in completing the teams, by the end of 2024 all personnel foreseen by the plan is in place and all formal deliverables have been completed. The shift in contract start dates and the long delivery times for hardware due to international market conditions have however prevented the project from fully spending the allocated budget in 2024, although a larger portion of the budget for an estimated 2.5M USD for personnel and 6M USD for material is already committed into 2025. The recruitment plan and the technical plans have been reprofiled to recover from the initial delays.

All Tasks are now actively working on the planned actions and have already produced notable results. All results and the ongoing activities have been presented at the [1st NextGen Technical Workshop](#) in November 2024.

## Introduction

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, 2025. 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 covers the activities performed during the calendar year 2024 and the status of milestones M1.0.1 through M1.4.2 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 2024 to January 2025). 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 2025. 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 described later in this document.

## NextGen Annual Milestones Status Sheet

Year	Code	Milestones	Type	Status
1	M1.0.1	Project management, risk management, activities and resources report	<a href="#">PM Report</a> <a href="#">Evaluation Report</a>	Ok
1	M1.1.1	Tender specification finalized and procurement launched for limited seeding resources.	<a href="#">Report</a>	Ok
1	M1.1.2	MLonFPGA community workshop, to identify needs from ATLAS and CMS on WP 2 and 3 as inputs to WP 1.2 and 1.3	<a href="#">Event</a> , <a href="#">Report</a>	Ok
1	M1.1.3	Workshop to identify and prioritize event generators' components suitable for acceleration, and develop LQFT benchmarking software tailored to hardware infrastructure	<a href="#">Event 1</a> , <a href="#">Event 2</a> , <a href="#">Event 3</a>	Ok
1	M1.1.4	Report on the preparatory work done in WP1.7 and the concrete work planned for year 2 of all of the sub-projects	<a href="#">Report</a>	Ok
1	M1.2.1	ML development toolkit for AI algorithms on FPGA at fixed latency for the ATLAS global trigger	<a href="#">Report</a>	Ok
1	M1.2.2	Initial algorithm for L0 Muon MDT with RPC and Tile seeding	<a href="#">Report</a> , <a href="#">Software</a>	Ok
1	M1.2.3	Review of identified physics scenarios for enhanced trigger	<a href="#">Report</a>	Ok
1	M1.3.1	Report on the online reconstruction performance, addressing identified bottlenecks, proposing specific improvement solutions, and outlining necessary features for the generic CMS Structure of Arrays (SoA)	<a href="#">Report</a>	Ok
1	M1.3.2	Report on the impact of RAW data compression and of their replacement with low-level reconstructed quantities (RAW')	<a href="#">Report</a>	Ok
1	M1.3.3	Completion of first studies on physics analysis prospects for Phase-2 L1 data scouting in simple final states, evaluating physics reach and requirements on the data scouting architecture (bandwidth, processing power)	<a href="#">Report</a>	Ok
1	M1.3.4	Implementation and integration of AI-based algorithms for the Run 3 L1 Global Trigger (including unsupervised algorithms for anomaly detection) and Phase-2 Correlator Trigger, including definition of operational practices for continuous training and deployment of models during Run 3	<a href="#">Report</a> <a href="#">Software</a>	Ok

1	M1.4.1	1st NextGen Triggers Project Workshop. Report on exchange and outreach activities	<a href="#">Event Report</a>	Ok
1	M1.4.2	CERN STEAM Programme governance and outreach channels in place	<a href="#">Web site Report</a>	Ok

A photograph of two people in business attire (a man in a grey shirt and a woman in a white shirt) standing at a wooden table. They are looking at and pointing to various business documents, including a line graph, a pie chart, and a bar chart. The woman is holding a pen. The scene is brightly lit, suggesting an office environment. A large, semi-transparent red 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 relations between CERN and Hillspire, external partners, the internal CERN services, and CERN and Experiments management. In December 2023, a Project Coordinator was appointed by CERN management to define the governance and implement the structure of the project. The Deputy and the rest of the Project Management team were fully in place by February 2024. The governance is described in the Deliverable M1.0.1 and includes inter alia 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.

In 2024 the main challenge was the implementation of the recruitment plan. This was mainly due to the fact that the closeness of the signature date of the agreement in November 2003 and the start date of the project in January 2024 did not allow recruiting in advance of the project start as foreseen. The majority of new positions had therefore to be opened during the year. Although by December 2024 all the planned positions have been fulfilled and all activities have started, many researchers and engineers started 3 to 6 months later than originally expected.

The original activity plan was already designed to take into account that recruitment would ramp up progressively during the course of 2024 and therefore all deliverables for 2024 have been successfully met as this report can demonstrate. However, a number of activities have started with some delay. In order to prevent the delays from impacting the work planned for 2025, the recruitment plan has been adapted to increase the personnel at the beginning of 2025 using the unspent budget from 2024 (see Fig. 1). At the moment of writing this ER and with the new recruitment and activity plan in place, we do not expect problems in recovering the delays in 2025.

The delays in recruitment and the fact that hardware procurement contracts tendered, negotiated and signed in 2024 will not result in actual costs until 2025 have caused a lower than expected consumption of the allocated budget. However, the shift in recruitment start dates and the signed procurement contracts represent a commitment of about 8.5M USD (2.5M USD for personnel and 6M USD for hardware and cloud services) that will be formally described in the 2025 reports bringing the expenditure plan in closer alignment to the expectations.

As far as the Project Management and Communications Work Package is concerned, all expected personnel are in place, including the Project Coordinator, a deputy, a

dedicated Project Manager, and the Communication Officer.

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



Fig. 1: Initial (top) Vs. revised (bottom) contracts start dates distribution (head count, in GREEN the contracts already awarded in 2024)

### **M1.0.1 Project management, risk management, activities and resources report**

The main deliverable from the Project Management and Communications work package is the publication of the Project Management Plan (inclusive of a Risk Management Plan and a Communications Plan) and the Year 1 Evaluation Report (this document) and Expenditure Responsibility Report.

The initial version of the Project Management Plan was developed during Q1 2024 and approved by the PMC and the SB to form the basis of the implementation of the project structure and governance. The Risk Management Plan and the Communications Plan have been developed in Q2 and are now in use as part of the regular project management activities. All necessary internal and external project management and communications tools are in place. The following proofs of evidence are provided:

#### **Deliverables**

Evaluation Report 2024 (this document)

Expenditure Responsibility Report 2024





**WP1:**

# **INFRASTRUCTURE, ALGORITHMS AND THEORY**

## Work Package 1: Infrastructure, Algorithms and Theory

### General Evaluation

All tasks (except Task 1.6, which is planned to start in 2025) have started in 2024. They have successfully completed the hiring phase, even if delayed compared to the initial scenario. The supervision and work items were adjusted as required by the slower ramp-up. The inclusion of all stakeholders posed a challenge to WP1: it is uncommon for CERN to involve technical experts from IT, TH, EP departments, and experiments at the same time. This makes WP1 an excellent opportunity for CERN to create new synergies. 2024 witnessed the power of this lateral integration, with several well-attended workshops involving multiple relevant parties.

To ensure integration of and communication between all stakeholders, shared task leaderships and regular meetings have been set up, both at the WP and the Task level, where applicable. 2024 has seen refinement in the assigned task leaders, as the leadership role transitioned from initial setup to coordination of technical work.

Recruitment was successful, filling all but 2 student positions foreseen for 2024. This additional person power made CERN a new focus point for several R&D areas such as hls4ml. Training (both formal and through their supervisors) and integration of the new hires have proven to be successful, with the new hires contributing to workshops such as the 1st Technical Workshop in November. Technical work is now ongoing throughout the project, using seeding resources made available by Task 1.1 while waiting for the ordered hardware to arrive. Neighboring R&D from within the high energy physics community has already started to align and coordinate with the related Tasks in WP1.

Thanks to this successful start-up phase and a change to productive technical work, as well as contributions from the community outside NextGen, all 2024 milestones have been achieved for WP1.

#### **M1.1.1 Tender specification finalized and procurement launched for limited seeding resources**

The first half of 2024 saw the completion of the [requirements collection](#) for NextGen project resources, covering hardware accelerators and other specialized deployments from the different tasks in the project.

The specification for all resource types was completed and the tender launched for the bulk of the requests, with successful bids selected. As of November 2024, the order for these bulk resources has been made with quote requests for the remaining resources - it is expected orders for all remaining resources will be out by mid-December 2024.

The effort covered the totality of resources for the NextGen project going beyond the initial goal of focusing only on seeding resources during the first year.

#### **M1.1.2 MLoFPGA community workshop, to identify needs from ATLAS and CMS on WP 2 and 3 as inputs to WP 1.2 and 1.3**

The "NextGen workshop: hls4ml HEP Community Forum" took place on 27th of September 2024. The goal of the workshop was to identify the needs from ATLAS (WP 2) and CMS (WP 3) as input to WP1 tasks 1.2 and 1.3. The agenda is linked [here](#). The agenda consisted of 12 presentations. The workshop started with two presentations by leads of WP1 tasks 1.2 and 1.3, giving an introduction to the NextGen project, setting the goals of the workshop, providing an overview of the current status of hls4ml and discussing future perspectives. This was followed by 3 presentations from ATLAS (WP2 2.1, 2.2 and 2.4), 3 presentations from CMS (WP3 3.5, 3.6 and 3.7), and 4 user contributions. Here are a few of the key takeaways from the workshop, summarizing the needs by the community, that highlight the direction of our research and development efforts for the coming years. There is a strong interest in exploring the computational capabilities of AMD/Xilinx AI engines for trigger applications. Therefore, support through hls4ml is highly desired. There is a keen interest in Graph Neural Networks (GNNs), DeepSets, Spiking Neural Networks (SNNs), and other innovative and "exotic" architectures. Most projects focus on low latency trigger applications. However, there are also studies on the use of FPGA accelerator cards in higher level trigger systems, which are not primarily latency constrained. To support this variety of applications with different constraints, an easy setup for hyperparameter optimizations is desired, which streamlines the process and enhances productivity. Enhanced reconfigurability, enabling dynamic adjustments to weights and architecture as needed, is a desired feature to improve operations in trigger systems. An hls4ml build server will facilitate the development and deployment of machine learning models on FPGAs. The workshop was a success, with experiments and the user community providing valuable feedback for tasks 1.2 and 1.3. This feedback will be incorporated into the upcoming deliverables with the updated plan to be presented at the NextGen technical workshop in late November.

#### **M1.1.3 Workshop to identify and prioritize event generators' components suitable for acceleration, and develop LQFT benchmarking software tailored to hardware infrastructure**

The workshop on [event generators' acceleration](#) was organized on 13-14 November, shortly before the formal start of the NextGen project, to ensure a prompt start of the activities and of the recruitment campaign. The outcome of the workshop discussions is now guiding the implementation of the task's goals; as a first concrete result, one of the key participants in the Workshop was hired on the NextGen staff position opened for this task.



[The workshop on Lattice QCD algorithms](#), organized on 9-11 December, brought together Lattice QCD experts and computing experts, to review challenges and opportunities offered by the novel hardware infrastructures, also in view of the resources being procured through the Nextgen grant.



Opening speech at NGT Lattice QCD algorithm workshop, held on December 9th to 11th.

A further [workshop](#) was organized on 4-5 November to stimulate community engagement around the theme of task 1.4. This triggered a new collaboration on the application of tensor network techniques to HEP, between the staff recruited for the NextGen project and experts from leading centers in Europe (Padova University - Italy; Wigner Centre - Hungary; Tech Univ Munich - Germany).

### **M1.1.4 Report on the preparatory work done in WP1.7 and the concrete work planned for year 2 of all of the sub-projects**

This task has stakeholders from all LHC experiments, as well as software experts from the physics and information technology departments. It is proving to be an extremely valuable venue, a rare occurrence of bringing the community's most renowned software experts together to work on key challenges for the whole community. This task has seen [measurable progress in its five R&D areas](#), despite the ramp-up turning out to be too optimistic due to hiring delays of about 6 months.

#### **Efficient heterogeneous scheduling**

An initial, simplified demonstrator based on the software framework used by two experiments has been created. It enables the researchers to implement and benchmark potential solutions, such as scheduling based on coroutines which is [currently being implemented](#). Coroutines address the challenge of algorithms having to yield to other tasks (for instance because they need to execute on the GPU) multiple times. Realistic algorithms will be implemented in 2025, to guide evaluation of different scheduling options.

#### **Efficient portable data structures**

Two first prototypes for the conversion between array-of-structs to struct-of-arrays have been implemented differently. The more speculative implementation uses C++ reflection; a more realistic implementation is based on traditional templates, building on experience from all four LHC experiments. Already at this early stage, both provide valuable expertise and training to the new personnel. In 2025, the implementations will be benchmarked; a possible third scenario combining the strengths of both current implementations will be benchmarked.

Furthermore, investigations on the usage of modern C++ data structures for heterogeneous executions, in the context of Monte Carlo event generators, have started.

#### **Common accelerated libraries**

A new post-doctoral researcher is expected to join this effort early 2025. Preliminary work on this subtask has started, for instance regarding operations of matrices with 20x20 elements. Nonetheless, much of the market research (to avoid re-implementation of existing solutions) and benchmark infrastructure will happen only in 2025.

#### **Efficient accelerator interfaces to ML inference**

A post-doctoral researcher and a doctoral student will join the task only in the beginning of 2025. First investigations have identified [Fermilab's SONIC](#) as a potential collaboration target.

### **Alternative programming languages**

A first study on Julia has been carried out, validating the simplicity and efficiency of the language. Several algorithms of [CMS Patatrack's](#) pixel reconstruction have been re-implemented in Julia by non-experts. Working on a single-threaded application, they were able to bring the performance of Julia-based implementation almost on-par with the C++-based implementation. An investigation on building shared libraries from Julia code is currently ongoing. In 2025, this sub-task will investigate other potential languages and criteria for determining their relevance, including interfacing C++ and accelerator support.





**WP2:**

**ENHANCING THE  
ATLAS TRIGGER  
AND DATA  
ACQUISITION**

## Work Package 2: Enhancing the ATLAS Trigger and Data Acquisition

### General Evaluation

First efforts in Work Package 2 (WP2) have been dedicated to re-evaluating our hiring plan in view of the new financial situation, rules and timescales adopted in the hiring of personnel. The adjusted plan preserved a mix of senior and long-term positions with student and early career positions. All 18 foreseen positions for 2024 were filled, with most of the people starting in June and July.

Task leaders ensured that the new hires could integrate well into the existing TDAQ and Offline tracking upgrade groups, to become effective rapidly and to leverage the existing expertise from these groups for their R&D work. The quality of the new hires and their strong commitment allowed us to successfully achieve the goals of the three contractual milestones for 2024, despite the initial delays in the project. An executive summary is described below, while more details may be found in additional documents available as references. Presentations at conferences and at ATLAS Internal Meeting may be found in a [dedicated archive](#).

### M1.2.1 ML development toolkit for AI algorithms on FPGA at fixed latency for the ATLAS global trigger

The ATLAS Global Trigger (L0Global) is part of the foreseen Level-0 System upgrade for the High-Luminosity LHC. It uses custom hardware to analyse Calorimeter and Muon Spectrometer data to make a decision to select or reject each of the events that ATLAS records every 25 ns. The Level-0 decision for each event has to be taken within a fixed latency of about 10  $\mu$ s. In order to achieve the required rejection of 40 in this event selection, L0Global aims at using close to offline reconstruction-like algorithms on the full data stream of 40Tbps coming from the detector.

The L0Global system aims at running all selection algorithms on single FPGA boards, deployed in a custom framework defining the datapath for data, control and timing signals, interconnected with about 100 multi-gigabit transceiver links per board.

This NextGen Triggers task aims to advance on the algorithm development, using novel ML techniques while staying within the boundaries of the existing hardware architecture.

During 2024 the following tasks were achieved, all part of the 12 months milestone and objectives:

- The initial goal achieved was to identify a set of suitable development and ML training tools. This toolkit provides the development framework with a defined



interface to HLS synthesis tools. The prototype framework has been exercised with a Boosted Decision Tree (BDT) dedicated to electron/photon selection using calorimeter data. The latency of this algorithm was measured and is well below the 10  $\mu$ s latency constraint. A first integration of a Convolutional Neural Network (CNN) workflow is underway and will be finalised by the beginning of next year.

- An important development during the last months has been the evaluation of new FPGA technologies that are becoming available in particular for the AMD ecosystem. The work has focused on the neural/tensor engines now available on production silicon, dedicated to scalar and vector processing, optimised for digital signal processing and Machine Learning. Preliminary results show that the use of the scalar processor to run the algorithm yields latencies well below the 10  $\mu$ s requirement.
- Next steps will be the assembly of a L0Global processor board, equipped with the new FPGA with AI engines. If these further tests are successful, the ATLAS collaboration may decide to adopt this new FPGA architecture, which includes the AI engines, as a new baseline for the L0Global system. This would be the first important achievement of the NextGen Triggers Project in ATLAS.

The more detailed 12 months milestone status report is available at the following [link](#).

### **M1.2.2 Initial algorithm for L0 Muon MDT with RPC and Tile seeding**

The ATLAS baseline Level-0 Muon trigger upgrade relies on fast trigger detectors, so-called Resistive-Plate Chambers (RPCs), in the Barrel region of the experiment. The goal of this task is to explore different algorithms and approaches that may increase the robustness of the Muon trigger in case of reduced coverage and performance of the RPC trigger detectors. Following the objectives of the 12 months milestones and objectives, this task work has concentrated on conceptual studies on both improving robustness and search for exotic signals:

- The approach followed has been to identify the processing steps for a trigger based on the Monitored Drift Tubes (MDT), using poorer seeds from the trigger detector and possibly using additional information from the Tile Calorimeter. In this scenario one needs to perform pattern recognition using MDT hits to find tracks. The timing information is taken directly from the MDT hits or from the additional Tile Calorimeter hit information.
- Preliminary studies on Machine Learning algorithmic approaches to Muon trigger processing have been carried out. A hybrid path has been developed for training data based on full simulation detector simulation with the new ATLAS Phase-2 Muon detector geometry and digitization, and the use of toy simulation for rapid algorithm development.  
The first approach has been to investigate CNNs for pattern recognition of muon tracks in presence of background hits. This study has also proven the

necessity of exploring algorithms suitable for sparse data such as RNNs and GNNs to further improve the performance.

- The need for optimal Machine Learning algorithmic techniques is even more evident for the search of new signatures, like displaced muons, where the standard algorithms available show poor efficiencies. The first study using a RNN has shown that this approach is good, also valid for the regression of physical quantities.
- Some advanced work has been possible, demonstrating the full prototype workflow which includes simulation, training, HLS code generation and RTL simulation. A new muon seeding logic has been implemented using only MDT hits, exploring the coincidence of hits at similar angular coordinates in the bending plane. The results seem to be compatible with timing constraints and the FPGA hardware resources available.

The deliverables of the work carried out during 2024 is available as a repository, at the following [links](#), while a report is linked [here](#).

### **M1.2.3 Review of identified physics scenarios for enhanced trigger**

The High Luminosity LHC Upgrade will allow ATLAS to collect an unprecedented amount of data to drive an exciting and diverse physics programme, from precision Higgs, QCD, top and electroweak physics, to searches for new physics. The scope of this task is to enhance the ATLAS online trigger event selection and in particular to boost its sensitivity for new physics. This is done through:

- Fully exploring the improved Level-0 and Event Filter reconstruction techniques developed in WP2, together with novel (end-to-end ML based) techniques for enhanced particle identification.
- Developing trigger concepts for exotic, non-standard signatures beyond the baseline Level-0 and Event Filter menus.
- Investigating Trigger-Level-Analysis (TLA), which allows to analyse all events recorded with ATLAS and to record a higher number of events with broader event selection criteria, but without retaining the full raw data payload.

During the course of 2024 the following tasks were achieved, to fulfill the 12 months Milestone and Objective:

- An extensive survey was prepared and sent to the ATLAS collaboration, in the context of the ATLAS TDAQ Physics Performance and Event Selection (PPES) group. A number of limitations in the current baseline trigger strategy have been identified in the following areas:
  - Enhanced Event Filter reconstruction for rare processes which are essential to characterise the self-coupling of the Higgs, a fundamental parameter appearing in the Higgs potential.

- A particular interest in non-standard signatures for exotic searches also shows the need for novel trigger concepts, not currently available in our baseline trigger selection strategy.
- Exploiting Trigger Level Analysis (TLA) will allow for Beyond the Standard Model and more exotic searches otherwise not accessible at the HL-LHC. The TLA idea is to store lightweight high-level information coming from trigger reconstructions (Event Filter jets, photons, muons, ...) for offline analysis. This work will leverage the experience obtained during current Run-3 trigger operations.

Based on the above studies, initial focus is on adding new Jet trigger chains to the HL-LHC trigger menu, while exotic triggers will follow soon. A dedicated event simulation for Di-Higgs scenarios with various BSM couplings and di-jet (background) events has been launched, taking into account different event pile-up conditions.

- A trigger emulation tool, that allows a fast evaluation of novel trigger selection strategies without a full trigger menu simulation, is also under development. The tool makes use of readily available offline reconstruction physics objects as an approximation for trigger reconstruction results that allow to quickly estimate rate and selection efficiency as a function of the trigger choices.
- A close collaboration with Work Package 1.6 has just started, to study signatures of new physics. The NextGen Triggers project will greatly enhance the opportunities for close collaboration between the experimental trigger and the theory communities.
- Lastly, more speculative and potentially groundbreaking R&D has also started, looking beyond the initial HL-LHC phase (Run-4). An opportunity might arise from the need to replace the inner two layers of the Phase-2 Pixel detector because of aging effects on the sensors due to the harsh radiation conditions at the HL-LHC. If 4D sensors with increased readout bandwidth could be used for the replacement, then novel end-to-end ML flavour tagging could potentially become available before the High-Level Trigger stage. A novel study in WP2 has started aiming at b-jet tagging using a Graph Neural Network/Transformer approach that explores the hit information of the replaced two innermost layers of the Pixel detector that provide timing to remove background from event pile-up.

Given the work done, we consider this task to be on-track, looking forward to finalizing baseline processes and key signatures, and being able to emulate trigger performance and estimate trigger rates. The trigger emulation framework will facilitate studies of dynamic allocation of trigger bandwidth, automating prescales and optimizing the ATLAS event selection which is stored for offline physics analysis. More details can be found in a [report](#).





**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.

The milestones M1.3.1 and M1.3.2 were critical components of the HLT-focused efforts within the NextGen project for CMS. These milestones shared a common objective: to evaluate the current capabilities of the CMS Phase-2 High-Level Trigger (HLT) systems and provide forward-looking projections to meet the ambitious goals of NextGen. Specifically, Milestone M1.3.1 concentrated on analyzing reconstruction performance, identifying bottlenecks, and proposing the speed-ups and data-structure improvements necessary to handle the increased data rates and complexities. Milestone M1.3.2 complemented this by exploring how event sizes could be reduced to make room for the additional foreseen data, through efficient compression techniques and the replacement of raw data with reconstructed quantities. Together, these efforts aimed to provide a clear understanding of where we stand today and what is required to meet the challenging NextGen goals. The detailed reports demonstrate the successful completion of these milestones, providing critical insights and solutions for the future of CMS Phase-2 HLT systems.

The milestones M1.3.3 and M1.3.4 were instead critical components of the L1T-focused efforts within the NextGen project for CMS. Specifically, Milestone M1.3.3 concentrated on providing a first prototype of the ambitious data-analysis facility that aims at analyzing all collision events using only the L1T event reconstruction, and on proving that the HL-LHC CMS physics program can be actually expanded by this approach. Milestone M1.3.4 is instead devoted to powering the upgrades of the L1 Data Scouting and Data Triggering workflows with AI. On one side the L1 Trigger reconstruction for HL-LHC is improved by ML algorithms for objects identification, on the other hand practises for unsupervised models are being developed for the first Anomaly Detection Trigger recently deployed in the current CMS data taking. Finally, developing and operating the experiment with a large amount of ML in the data acquisition pipeline is a new frontier for CMS. The L1T NextGen team has kickstarted a big effort towards automating all the operations related to training and deploying models. The achievements obtained throughout this first year pave the way for tackling the upcoming milestones with confidence; the detailed reports below demonstrate the successful completions of M1.3.3 and M1.3.4.

### **M1.3.1 Report on the online reconstruction performance, addressing identified bottlenecks, proposing specific improvement solutions, and outlining necessary features for the generic CMS Structure of Arrays (SoA)**

The [document](#) represents the detailed report required to fulfill the contractual milestone for 2024, demonstrating significant progress within the NextGen Task 3.1 framework. It comprehensively evaluates the performance and optimization strategies for the high-level trigger (HLT) and offline reconstruction systems, essential for meeting the demanding requirements of NextGen operations at the High Luminosity Large Hadron Collider (HL-LHC). The report thoroughly examines computational efficiency and resource usage under simulated high pile-up conditions (140PU and 200PU), highlighting key bottlenecks in resource-intensive modules such as `RecoTracker`, `RecoLocalCalo`, and `RecoHGCAL`. Leveraging recent CMSSW releases, it provides detailed benchmarking results and extrapolations to project the required performance improvements, accounting for anticipated advances in computing hardware and algorithmic refinements. The analysis identifies specific developments needed, such as hardware acceleration, GPU utilization, and novel data structures, to enable offline-like reconstruction quality at a data rate of 750 kHz. This work not only confirms the team's ability to deliver the detailed technical report as per the 2024 milestone but also lays a solid foundation for continued progress in optimizing reconstruction systems for HL-LHC's unprecedented data demands.

### **M1.3.2 Report on the impact of RAW data compression and of their replacement with low-level reconstructed quantities (RAW')**

In the first year, the Task 3.3 team initiated efforts to reduce CMS event sizes, aiming to enable higher output trigger rates. After a few months spent selecting a suitable doctoral candidate, the chosen student began their research activities. This task also benefited from collaboration with the PRIN2002 project, "PRE: Partially Reconstructed Event."

The primary objective was to measure the RAW event size of high-pileup Run-3 proton-proton collision data and evaluate the impact of the RAW' compression method, currently employed in Hlon collisions. This method replaces raw data from the silicon strip tracker with reconstructed strip clusters, using compact representations such as average charge and barycenter instead of raw ADC data. The silicon strip tracker, identified as the largest contributor to raw event size (55%), experienced a significant reduction in size—about 20%—when the RAW' technique was applied. Preliminary tests indicate that optimizing the number of bits for specific

variables could enhance compression further, achieving reductions of up to 30%.

Parallel studies were conducted on Phase-2 upgrades, particularly focusing on raw detector data sizes. Since the RAW data format for Phase-2 is yet to be finalized, size estimates were based on simulated detector digitization (simDigis). The NextGen team found that, for Phase-2 simulations with pileup 200, the high-granularity calorimeter (HGCAL) and the inner tracker will be the largest contributors to data sizes, at 2.1 MB/event and 1.4 MB/event, respectively. In contrast to Run-3, the Phase-2 outer tracker will account for only 15% of the event size, as its data will be stored in a simplified single-bit (0 or 1) format instead of using ADC counts.

Additionally, the team analyzed the sizes of low-level reconstructed variables, such as particle positions and energies within the detector (rechits). Their findings revealed that replacing the inner tracker raw data with rechits could reduce its size by up to 89%, while replacing HGCAL raw data with rechits could achieve a reduction of up to 59%. All findings have been summarized in a report available at this [link](#).

### **M1.3.3 Completion of first studies on physics analysis prospects for Phase-2 L1 data scouting in simple final states, evaluating physics reach and requirements on the data scouting architecture (bandwidth, processing power)**

A first set of prototype physics analyses have been studied, covering searches for rare decays of the W and Higgs bosons and the Bs meson. For all the physics analyses studied, the performance of the event reconstruction performed in the L1 trigger and scouting system is found to be satisfactory: the invariant mass resolution is sufficient to reveal a potential signal as a narrow resonant peak over the background, and a good selection efficiency and background rejection can be achieved already with simple selection criteria. First projections for one year of data taking at HL-LHC yield promising physics sensitivity.

A demonstration system has been set up to test data acquisition and real-time analysis for L1 Scouting, with prototype CMS L1 Trigger boards, FPGA development kits and a small cluster of 5 servers connected via high-speed optical links and network switches. The NextGen project has been essential in the development of this demonstration system, as all the hardware components other than the L1 Trigger boards were procured using NextGen funding, and the system was set up and tested by NextGen personnel (students and supervisors). Data acquisition of streams of inclusive particles reconstructed in the L1 Trigger was successfully demonstrated as well as data decoding and live processing running all the nine W and Higgs boson prototype analyses was also demonstrated for an event rate of 27 MHz and input data rate of about 12 GB/s.

All supporting materials detailing and demonstrating the outcomes achieved are provided in the dedicated [M1.3.3 Report](#), which includes links to all presentations and documents produced throughout the year.

### **M1.3.4 Implementation and integration of AI-based algorithms for the Run 3 L1 Global Trigger (including unsupervised algorithms for anomaly detection) and Phase-2 Correlator Trigger, including definition of operational practices for continuous training and deployment of models during Run 3**

Thanks to new personnel dedicated to the NextGen 3.6 Task, significant progress has been made on implementing and integrating AI-based algorithms for the HL-LHC L1T Correlator Trigger, particularly on flavor tagging and calorimeter cluster identification. With flavour tagging, we aim to identify the flavour of the original particle produced in the collision that initiated a jet. Building on R&D into efficient Neural Network architectures on FPGAs, a new approach investigated this year uses a permutation-invariant DeepSets architecture. This Neural Network architecture derives information from each constituent particle of the jet, and aggregates across the set of constituents before making a final classification. Before jet-clustering, the constituents of the jet have to be built from calorimeter clusters and tracker tracks. Another important achievement obtained during the year is the improvement in the identification of the calorimeter clusters sent to the High Granularity Calorimeters to the L1T, with a multiclass BDT, which simultaneously separates the clusters into  $e/\gamma$ , pion, or pileup. The improvements provided by this new approach are especially pronounced for low transverse momentum electrons, which are of particular interest for the Phase-2 L1 Scouting system.

Regarding the current data taking, a first-ever real-time anomaly detection algorithm (AXOL1TL) that leverages unsupervised learning and features an autoencoder that operates within 50 ns in the L1 Global Trigger, has been deployed in Spring 2024. Since then, the personnel hired with NextGen 3.7 Task, who joined the AXOL1TL team, started working on the unprecedented challenges related to this novel triggering technique. The team is investigating the collected anomalous event data for potential new physics signals, automating the operations related to training and deploying models, designing a more robust model based on representation learning techniques, and developing an upgraded model tailored to the particle level inputs and architecture of the Phase-2 L1 trigger system. The outcomes of these activities are critical for leveraging advanced ML to maximize the discovery potential of the LHC's Run 3 and the HL-LHC datasets.

The dedicated [M1.3.4 Report](#) provides all supporting materials detailing and demonstrating the outcomes achieved. It also includes links to presentations, documents, and software produced throughout the year.





**WP4:**

**EDUCATION  
PROGRAMMES  
AND OUTREACH**

## Work Package 4: Education Programmes and Outreach

### General Evaluation

Work package 4 focuses on ensuring the development of world-class high-energy scientists and engineers who are inside the NextGen project or working outside in High Energy Physics experiments at CERN or in other research laboratories worldwide.

The activities are organized in two sub work packages in different areas (Outreach and Education) with complementary goals but both in close collaboration with academic and industry partners to ensure future growth in this field, dissemination, and the sustainability of the results obtained.

The start of activities on the Outreach task 4.1 was delayed by approximately 6 months due to the hiring ramp-up. The Education task 4.2 was able to start immediately and has focussed on ensuring proper training was available to the scientists that were newcomers to the NextGen project. This is detailed in the two following sections.

### M1.4.1 1st NextGen Triggers Project Workshop. Report on exchange and outreach activities

#### Outreach activities

The first step in the outreach efforts for NextGen Triggers project was to establish a Communications Plan (as described in the Project Management section in this document) and dedicated communication channels. To achieve this, the [NextGen website](#) was designed and deployed at CERN, which serves as a central hub where different audiences can access all relevant information about the project activities. This platform provides updates, resources, and insights into the development and impact of NextGen Triggers, aiming to make this topic accessible to a broader audience.

In addition to this, throughout the year, NextGen Triggers has organized and collaborated on different events, the main one being the three-day [1st Technical Workshop](#) in November 2024. For all these events, two main channels have been primarily used: the “News and Events” section of the website and the LinkedIn channel. Additionally, we have also leveraged other platforms, such as the CERN main website’s “news” channel and the “IT Read Me” blog, to further extend our outreach and engage with a broader audience. Articles were published in various conferences and on the CERN Courier publication. The project was also present on the CERN social media channels.



The CERN NextGen Triggers 1st Technical Workshop attendants, November 2024.

## The Exchange programme

The education and outreach work package has also a funded exchange programme aiming to enable world-class scientists and engineers to collaborate with academic and industry partners.

The goal is to ensure continued skills development of world-class scientists and engineers able to combine domain-specific knowledge of high-energy physics with data science and artificial intelligence.

The programme expects to have these exchanges taking place in the technical work packages 1, 2, and 3, which have all expressed interest in participating in this activity.

The exchange will be organized by allowing scientists and researchers to come to CERN to facilitate the knowledge sharing with the project experts and promote and encourage project members to visit external institutes and companies.

Due to the hiring delays that the project experienced, we have focussed on ensuring the programme will start in 2025 by clarifying what would be the administrative framework that would allow external visitors to receive subsistence allowances and/or cost of living allowances.

The detailed plan is being prepared and work is ongoing in the technical work packages to establish links with academia, research laboratories, and industrial partners.



### **M1.4.2 CERN STEAM Programme governance and outreach channels in place**

During the first year of the STEAM initiative, foundational steps were taken to equip researchers with advanced software skills essential for high-energy physics research. Two committees were established: one dedicated to developing and implementing educational activities, and another focused on reviewing and refining the curriculum and its delivery methods.

A comprehensive skills gap analysis was conducted using surveys and interviews with task leaders, newly hired students, and graduates. This process highlighted critical training needs in high-throughput and real-time computing, data science methods and tools, and advanced neural networks. Engagement with established international schools—such as the CERN School of Computing, INFN Efficient Scientific Computing, and the SMARTHEP FastML School—resulted in about 20 individuals from NextGen Triggers receiving specialized training tailored to their requirements, including advanced C++ programming, machine learning, heterogeneous computing, and real-time inference techniques.

A central Learning Hub was created on the NextGen Triggers website, providing access to recorded lectures, seminars, and open-source tutorials, including a featured tutorial on the Alpaka performance portability library. Feedback collected from participants will guide the development of a structured curriculum in the coming year. This evolving framework will serve as a reference for future educational initiatives, paving the way for the STEAM program's official implementation phase scheduled to begin in 2026.

## Conclusions and Recommendations

The first year of activities of the NextGen Triggers project was mainly dedicated to setting up the project structure and governance and form the technical teams recruiting experts for the different tasks. Due to the very short lead time between signature of the Grant Agreement and start of the project, the recruitment took place during the execution of the project introducing some delays in using the allocated budget. A reprofile of the workforce and activity was designed to compensate for the delays starting already in the first quarter of 2025.

Thanks to the collective effort of the experts in the ATLAS and CMS experiments and the CERN Departments involved in the project, the planned milestones and deliverables were nonetheless achieved. The project is now in full operational mode and a number of documented results in assessing new technologies and methodologies have been developed and presented at events across the community.

Procurement contracts for the computing resources needed by the activities have been issued following the CERN standard procurement process. Due to the administrative steps required by the process and by market constraints, the time necessary to commission the full resource infrastructure is expected to take until mid-2025. This was already factored into the project plans and appropriate mitigations have been put in place (e.g. use of existing computing resources at CERN).

One of the most important lessons learned is that setting up a project of the size and ambition of NextGen Triggers and especially the time needed to identify and recruit experts of the appropriate level requires a dedicated process and sufficient lead time and flexibility.

## Appendix 1

## Milestones for the Coming Year 2025

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

Year	Code	Milestones	Type
2	M2.0.1	Project management, risk management, activities and resources report	Report
2	M2.1.1	Purchase of hardware and services and commissioning completed for on-premise and cloud resources.	Report
2	M2.1.2	hls4ml software release 1 with open-access documentation	Software
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	Report
2	M2.1.4	Workshop at CERN for discussing the status and plans for all of the sub-projects in WP1.7.	Event, Report
2	M2.2.1	First integration of ML in L0 Global trigger for commissioning and preparation of further improvements.	Software
2	M2.2.2	Prototype GNN tracking algorithm based on ACTS	Software, demo
2	M2.2.3	Prototype ML and ACTS based muon reconstruction algorithm	Software, demo
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
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

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
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.	Report
2	M2.4.1	2nd NextGen Triggers Project Workshop. Report on exchange and outreach activities	Event, report
2	M2.4.2	Skills gaps analysis done. Report on first year of the STEAM Programme activities	Report