Next Generation HEP Triggers Proposal

CERN - European Organization for Nuclear Research

V 1.2_dist (20231212)

Executive Summary

The High Energy Physics (HEP) program at CERN has achieved major breakthroughs in particle physics, technology, and algorithms, including the discovery of the Higgs boson in 2012. This allowed the validation of compatibility of the theoretical construction behind the Standard Model (SM) of particle physics with the data, but the existing uncertainties leave room for models beyond the SM. With the experimental collider framework in place, scientific exploration continues to answer questions around the origin of dark matter, the disproportionately low abundance of antimatter and the nature of the discovered Higgs boson. Hard physics problems aside, much can be gained from improvements to the data acquisition pipeline allowing for capturing a richer set of collision events, furthering scientific understanding.

The Large Hadron Collider (LHC) consists of a 27 km tunnel where superconducting magnets guide bunches of protons, circulating in opposite directions, which are then caused to collide at experimental sites (e.g. ATLAS and CMS) at a rate of 40 million times per second. The collision events emit various particles, which are tracked through a multitude of radiation-hardened detectors and fed into the L1 trigger system, which needs to reject >99% of the events within 10 microseconds due to detector cache constraints and available network capacity.

This data is further reduced by >99% in the High-Level Trigger (HLT) to conform to the current event analysis and simulation capacity. HEP experimentation is fundamentally stochastic, so without changing other factors, an increase in data collection throughput would allow for higher confidence in current results while increasing the likelihood of detecting novel particles in the current LHC setup. Furthermore, this capacity increase is absolutely needed for future LHC upgrades where each collision will have many more interesting events.

The interpretation of the LHC data relies on theoretical simulations of particle interactions in the Standard Model (SM) and in scenarios of new physics beyond the SM (BSM). The full exploitation of the immense HL-LHC datasets, and in perspective of the data from Future Colliders, will require radical improvements in the computing strategies of theory calculations, to increase their accuracy while keeping affordable computing times. A multitude of theoretical tools must be addressed, in a coordinated effort, to preserve their interoperability and harmonize the overall precision. In addition to the several ingredients needed to describe the final states of proton collisions, the infrastructure developed for the triggers, e.g. the GPU cluster, also supports the advancement of software and algorithms for lattice Quantum Field Theory (LQFT) calculations, as a unique approach to control relevant non-perturbative ingredients. The engagement of LQFT experts would also bring to the trigger community complementary expertise and experience in parallel architectures. The progress envisaged with these theoretical tasks complements and augments the benefits of the increased capacity to trigger and record relevant data.

The goal of this proposal is to facilitate improvements to LHC data collection and processing beyond current capabilities, while looking forward to future data collection needs, through four work packages. The R&D work done to optimize the current Run 3 and the following High-Luminosity (HL)-LHC phases will provide critical insight to develop future detectors and data flows for the even more ambitious objectives of the Future Circular Collider (FCC) currently in its Feasibility Study phase. We consider that such an ambitious programme requires co-development partnerships with experts in academia and industry to accelerate the achievement of the objectives

CERN Open Science Policies

Openness is a key value and principle that has been enshrined in the CERN founding convention for almost seventy years and was reaffirmed in the update of the European Strategy for Particle Physics in 2020. From the ultimate recognition of the universal importance of the fundamental scientific knowledge produced at CERN, the Organization derives the duty to make this knowledge available to everybody and the key role of open science in the pursuit of CERN’s mission. Supported by long-term financial investment from its Member and Associate Member States, with significant contributions also from non-Member States, CERN is committed to the advancement of science and the wide dissemination of knowledge by embracing and promoting practices making scientific research more open, collaborative, and responsive to societal changes.

To better align and coordinate the numerous open science activities across all CERN departments, and to strengthen its universal commitment to openness, the CERN Open Science Policy[1] was published on October 1st 2022. It reflects the Organization’s aim to be recognized as a leader in the open science domain. Concrete implementation measures are summarized in a dedicated Open Science Policy Implementation Plan[2]. The CERN Open Science Policy builds on the previously released CERN Open Access Policy[3] (last revision 2021), the CERH LHC Open Data Policy[4] (issued in 2020), as well as numerous decentral open science initiatives across all open science domains, for example the Open Source License Task Force that was initiated back in 2012.

To enable the efficient application of open science principles at CERN and beyond, CERN is partnering with other research institutions as well as funding agencies in developing open and inclusive repositories: The CERN Document Server (CDS) is CERN's primary institutional repository to allow easy dissemination of all types of CERN publications. The CERN Open Data Portal was specifically designed to allow the release of large and complex LHC datasets, the CERN Analysis Preservation tool helps data scientists in capturing all relevant information along a scientific analysis to enrich future data releases enabling easy reproducibility, and INSPIRE aggregates all types of research artefacts (articles, seminars,

[1] https://cds.cern.ch/record/2835057/files/CERN-OPEN-2022-013.pdf
[2] https://cds.cern.ch/record/2856044/files/CERN-OPEN-2023-007.pdf
[3] https://cds.cern.ch/record/1955574/files/CERN-OPEN-2021-009.pdf
[4] https://cds.cern.ch/record/2745133/files/CERN-OPEN-2020-013.pdf

datasets) enabling easy search and cross-referencing across the works. CERN also operates Zenodo, a HEP-inspired repository open for all researchers in the world from diverse disciplines that want to easily and openly share their research artefacts.

The CERN Open Science Policy framework is deeply embedded in the CERN institutional governance: the Open Science Steering Board, reporting to the Director for Research and Computing, is overall responsible and accountable for the institutional open science strategy and its alignment with general institutional strategic goals. The Board reviews and updates the open science policies and ensures external communication through a biennial CERN Open Science report (first issue to be published in 2024). Individual open science subject expert groups and subcommittees are tasked with the actual implementation of the Open Science policy. For this proposal, the most relevant bodies are the CERN LHC Open Data Working Group and the CERN Open Source Programme Office (OSPO), which was formally approved and is now in the process of formation. All Open Science activities are centrally coordinated and monitored by the CERN Open Science Office, which is part of the CERN Scientific Information Service, a cross-departmental service unit reporting to the Director for Research and Computing.

CERN pledges to release all IP generated as part of the NextGen Triggers project under appropriate open licenses in compliance with the CERN Open Science Policy, as described above.

Project structure and main milestones

WP1: *"Infrastructure, Algorithms and Theory" to improve ML-assisted simulation and data collection, develop common frameworks and tools, and better leverage available and new computing infrastructures and platforms.*
WP2: *"Enhancing the ATLAS Trigger and Data Acquisition" to focus on improved and accelerated filtering and exotic signature detection.*
WP3: *"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.*
WP4: *"Education Programmes and Outreach" to foster and train computing skills in the next generation of high energy physicists.*

Year	Description[5]	Type
1	15 nodes: 4x A100 80GB GPU (NVLink), 64 CPU, 512GB RAM	On-Prem
1	200k hours A100 80GB GPU	Cloud/HPC
1	6 nodes with 64 CPU, 512GB RAM, high speed NIC + 10x GPU/FPGA accelerators (WP3.1)	On-Prem

[5] The actual type of hardware (GPU type, FPGAs, ASICS, or other specialised hardware) and overall characteristics might change during the course of project depending on needs, benchmark results, and evolution of the products

1	Prototype CMS L1 hardware (WP3.5-3.7)	On-Prem
2	75 nodes: 4x GPUs, 64 CPU, 512GB RAM	On-Prem
2	Prototype ATLAS L0 hardware (WP2.1/2.2)	On-Prem
2	FPGA/SoC (WP2.3)	On-Prem
2	Prototype CMS L1 DAQ board components + PCB + assembly	On-Prem
2	External GPU + High Level ML Services	Cloud/HPC
3	External GPU + High Level ML Services	Cloud/HPC
3	2 node: 8x GPU/FPGAs, 64 CPU, 512GB RAM, High performance NIC (WP3.1)	On-Prem
3	Updated prototype CMS L1 hardware (WP3.5-3.7)	On-Prem
4	Updated prototype CMS L1 hardware (WP3.5-3.7)	On-Prem
4	External GPU + High Level ML Services	Cloud/HPC
5	External GPU + High Level ML Services	Cloud/HPC

Estimated distribution, type and costs of computing resources

Year	Code	Milestones	Type
1	M1.0.1	Project management, risk management, activities and resources report	Report
1	M1.1.1	Tender specification finalized and procurement launched for limited seeding resources.	Report
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	Event, report
1	M1.1.3	Workshop to identify and prioritize event generators' components suitable for acceleration, and develop LQFT benchmarking software tailored to hardware infrastructure	Event
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	Report
1	M1.2.1	ML development toolkit for AI algorithms on FPGA at fixed latency for the ATLAS global trigger	Report
1	M1.2.2	Initial algorithm for L0 Muon MDT with RPC and Tile seeding	Software

1	M1.2.3	Review of identified physics scenarios for enhanced trigger	Report
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)	Report
1	M1.3.2	Report on the impact of RAW data compression and of their replacement with low-level reconstructed quantities (RAW')	Report
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)	Report
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	Report, Software
1	M1.4.1	1st NextGen Triggers Project Workshop. Report on exchange and outreach activities	Event, report
1	M1.4.2	CERN STEAM Programme governance and outreach channels in place	Web site
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 techniques.	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
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
4	M4.0.1	Project management, risk management, activities and resources report	Report
4	M4.1.1	Report on hardware, service and platform efficiency for the different use cases. Demonstrator of full integration with CERN IT services and experiment workflows.	Report, Demo
4	M4.1.2	hls4ml and NNLO software release 2 with open-access documentation	Software
4	M4.1.4	Report on new triggers' performance on new physics benchmarks. Publish a study of the quantum machine learning foundations: from model building to efficient optimizers	Report
4	M4.1.4	Complete the development of the task scheduler prototypes. Produce a recommendation for the experiments on heterogeneous task scheduling.	Report
4	M4.2.1	Extended Global trigger algorithms ready for integration	Software
4	M4.2.2	Improved muon L0 trigger algorithm ready for its integration	Software

4	M4.2.3	Implementation of trigger signatures for identified physics scenarios	Software
4	M4.3.1	Implementation of a full-scale prototype that buffers the entirety of the RAW data for about 8 hours, derives improved calibrations, and uses them for the online reconstruction, event selection, and RAW’ data reduction.	Demo
4	M4.3.2	Implementation of a client-server, multithreaded, distributed test application, based on the CMS software framework CMSSW, leveraging high-speed host-to-host or shared memory communication.	Software
4	M4.3.3	First prototype of next generation ATCA data acquisition board for Phase-2 L1 Scouting produced and tested, architecture for extended Phase-2 L1 Scouting demonstration system defined, and started procurement of components (e.g. higher speed networking equipment, new accelerator cards, additional servers, …)	Demo
4	M4.3.4	Successful co-training of AI algorithms for different subsystems of Phase-2 L1 Trigger: physics performance documented in CMS public notes & conference talks/proceedings, firmware and emulators completed and integrated, training and deployment operational practices defined and documented.	Report
4	M4.4.1	4th NextGen Triggers Project Workshop. Report on exchange and outreach activities	Event, report
4	M4.4.2	Report on second year of full STEAM Programme activities	Report
5	M5.0.1	Project management, risk management, activities and resources report	Report
5	M5.1.1	Completed integration of project use cases with the hardware, services and platform chosen. Report on project results, open items and sustainability.	Report
5	M5.1.2	hls4ml and NNLO software release 3 with open-access documentation	Software
5	M5.1.3	Document first results of application of LQFT to spectral-reconstruction problems. Prove readiness for event mass production of LHC events on GPUs	Report
5	M5.1.4	Present the results of all WP1.7 sub-projects at major computing conferences.	Report
5	M5.2.1	Complete enhanced L0 Global and L0 Muon trigger strategies implemented	Report

5	M5.2.2	Enhanced ACTS based overall and muon tracking software fully integrated	Report, Software
5	M5.2.3	Enhanced high throughput data-collection fully optimized and integrated	Report
5	M5.3.1	The complete HLT R3 reconstruction is integrated in CMSSW, and its performance is compared with that of the baseline online and offline reconstruction.	Report
5	M5.3.2	Efficient data-oriented data structures are integrated in CMSSW, used by the R3 reconstruction algorithms running on heterogeneous hardware, and interfaced with ROOT-based persistent storage.	Report
5	M5.3.3	Extended demonstration system for Phase-2 L1 Scouting with input bandwidth >1 Tbps and capable of running the analyses identified in the physics studies of year 3 individually. Implementation and results documented as CMS public notes & conference talks/proceedings.	Demo
5	M5.3.4	Completed recasting of Run 3 L1 anomaly detection to Phase 2 particle-based inputs, and L1 scouting data compression, including firmware implementation and validation of physics performance. Implementation and results documented as CMS public notes & conference talks/proceedings.	Report, Software
5	M5.4.1	5th NextGen Triggers Project Workshop. Report on exchange and outreach activities	Event, report
5	M5.4.2	Report on third year of full STEAM Programme activities. STEAM Programme sustainability proposal ready	Report

Project milestones per WP and year (the hardware costs are attached to the milestones of the year during which they are purchased although they are used also in following years)

Work Packages

Project management and communications

The NextGen Trigger project mobilises considerable amounts of resources across CERN and external research and commercial partners. It is critical to the success of the project to dedicate effort to the overall project coordination, the management of the relations between CERN and the Hillspire Foundation, the external partners, and the internal CERN services. In addition, given the complexity of the programme and the expected visible impact on CERN experiments and beyond HEP in terms of technology and results, dedicated effort is required to manage the financial and reporting tasks and the communications activities.

Work Package 1: Infrastructure, Algorithms and Theory

The CERN teams from ATLAS, CMS, IT and theory (TH) will be developing cutting-edge AI and physics simulation algorithms and common applications across multiple experiments. Network development and optimization will enable heavy computing tasks, e.g., extreme-scaling simulations, training and Neural Architecture Search (NAS). Physicists and computer scientists in TH, CMS, ATLAS and IT will also develop tree tensor networks, classical and quantum algorithms for Lattice-quantum-field theory (LQFT) simulations of increasing complexity. We will use a local dedicated cluster of O(100) low-latency/high-bandwidth interconnected GPUs of the latest generation, arranged similarly to what is done at HPC supercomputer centers, and access to commercial cloud platforms and resources, both classical and quantum.

The associated increase in heterogeneous low-latency/high-bandwidth parallel computing resources will also require matched software and algorithm development, and an increase in the data network interconnects and datacenter capacity. The improvements to the computational infrastructure, and resulting gains in algorithm performance, are intended to improve the robustness of the experimental data collection, the scope of physics simulations, and reconstruction, leading to increased experimental efficiency and predictivity of theory simulations.

Work Package 2: Enhancing the ATLAS Trigger and Data Acquisition

This work package focuses on an enhancement of the already ambitious upgrade of the ATLAS experiment’s trigger and data acquisition system for the High Luminosity Phase of the LHC (HL-LHC) scheduled to start in 2029. Novel approaches to trigger event selection will be developed that will extend the ATLAS physics potential, in particular exploiting state-of-the-art Machine Learning techniques. Efficient use of acceleration technologies will be investigated to extend the capabilities of the trigger and data acquisition system, while improving the system’s energy efficiency. A successful implementation of this work package will result in the collection of richer collision events, extend the experiment’s sensitivity to a broader range of new physics beyond the Standard Model scenarios, and enable ATLAS to exploit state-of-the-art processing architectures in its online event selection to achieve the best possible physics performance at HL-LHC.

Work Package 3: Rethinking the CMS Real Time Data Processing

This working package aims to rethink the CMS data acquisition system allowing the CMS physics program to operate over all the collisions produced by the LHC. This is achieved through the High-Level Trigger (HLT) Real-time Reconstruction Revolution (R^3) and a novel L1-trigger scouting stream. These goals are achieved by mixing traditional physics-inspired algorithms with cutting-edge AI solutions and leveraging synergy between CMS physicists and data scientists. Proof-of-concept R&D projects on these activities are ongoing. Some of them already established the validity of the ideas behind the proposed tasks.

The planned tasks are independent efforts that will be carried on in parallel across a five-year period. The proposed work will deliver intermediate milestones on a two-year time scale that could be deployed in the CMS HLT during Run3, while parallel work streams will focus on developing the L1 trigger for Run3, with the possibility of deploying next-generation algorithms on a five-year time scale.

Work Package 4: Education Programmes and Outreach

The education and outreach work package aims to enable exchanges and 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 proficiency. This will be done in close collaboration with academic and industry partners to ensure the knowledge is both relevant and up-to-date. The activities are organized over two different areas with different but complementary goals: (i) promoting exchanges by allowing scientists and researchers to come to CERN and work with the project experts and project members to visit external institutes and companies; organise thematic events, project meetings, and outreach activities; (ii) designing and prototyping the CERN STEAM programme, a focused set of complementary activities, courses, training opportunities on AI and Data Science for HEP, providing a specialization path for researchers and engineering working on advanced computing for fundamental science.

Technical Annex

Detailed description of Work Packages, Tasks and Activities

Project management and communications

Task 0.1: Overall project coordination, partnership management, finance and communications

The NextGen Trigger project mobilises considerable amounts of resources across CERN and external research and commercial partners. It is critical to the success of the project to dedicate effort to the overall project coordination, the management of the relations between CERN and the Hillspire Foundation, the external partners, and the internal CERN services. In addition, given the complexity of the programme and the expected visible impact on CERN experiments and beyond HEP in terms of technology and results, dedicated effort is required to manage the financial and reporting tasks and the communications activities.

Time	Description	Deliverable/Milestone
6 m	Set up of project governance, internal and external communications channels, risk analysis and management, reporting mechanisms, etc	Governance is in place, communications and reporting channels are available and published, risk register in place
12 m	Project management, risk management and mitigation, communications, education, and reporting	Project is on track, risk register is regularly reviewed by the project manager and WP leaders and mitigation actions performed, project activities and results are broadly communicated via CERN and other channels, reports on activities and use of resources are published
24 m		
36 m		
48 m		
60 m		

Work Package 1: Infrastructure, Algorithms and Theory

The CERN teams from ATLAS, CMS, IT and theory (TH) will be developing cutting-edge AI and physics simulation algorithms and common applications across multiple experiments. Network development and optimization will enable heavy computing tasks, e.g., extreme-scaling simulations, training and Neural Architecture Search (NAS). Physicists and computer scientists in TH, CMS, ATLAS and IT will also develop tree tensor networks, classical and quantum algorithms for Lattice-quantum-field theory (LQFT) simulations of increasing complexity. We will use a local dedicated cluster of O(100) low-latency/high-bandwidth interconnected GPUs of the latest generation, arranged similarly to what is done at HPC supercomputer centers, and access to commercial cloud platforms and resources, both classical and quantum.

The associated increase in heterogeneous low-latency/high-bandwidth parallel computing resources will also require matched software and algorithm development, and an increase in the data network interconnects and datacenter capacity. The improvements to the computational infrastructure, and resulting gains in algorithm performance, are intended to improve the robustness of the experimental data collection, the scope of physics simulations, and reconstruction, leading to increased experimental efficiency and predictivity of theory simulations.

A range of specific interdisciplinary tasks connecting IT, theory and experiment have been identified.

Task 1.1: Procurement of hardware and services for large scale NN optimisation and training, and physics simulation

This task will focus on designing, procuring, deploying and operating the computing infrastructure (hardware and software) and platforms required to support the common tasks in WP1 (hardware-aware neural network training workflows and next-generation physics simulations) and the specific activities in WP2 and WP3. To make sure activities can start as early as possible, an initial amount of cloud-based resources will be procured as a way of benchmarking different architectures and algorithms. Once a better understanding of the requirements is established, hardware resources will be procured to be installed in the new CERN Data Center and dedicated to the task of this project. Dedicated effort is required for designing the specifications, supporting the procurement process, commissioning the resources, deploying and maintaining the software tools and frameworks, and monitoring exclusive use of the resources by researchers of this project.

Time	Description	Deliverable/Milestone
6 m	Overall system design/specification and estimation of initial infrastructure requirements, incl. resource layout across on-premises and public cloud service providers. Preparation of procurement specifications. Provisioning of limited seeding resources for setup tasks in WP1/WP2/WP3.	Tender specification ready for procurement. Testbeds are available and accessible with limited seeding resources.
12 m	Initial market survey and providers identification, execution of CERN procurement process. Research and development of a common MLOps platform for automation of the different steps of distributed training, AutoML and inference building on industry standards.	Approval of the first tendering process in place. Comparison report on potential platforms supporting the project use cases.

Time	Description	Deliverable/Milestone
12 m	Identify reference use cases for the whole project targeting distributed training, AutoML and inference. Validation of end-to-end workflows (training, optimization, serving) from project use cases.	Recommendation of best platform and tools for efficient usage by the project reference use cases.
18 m	Purchase of hardware and services based on successful tenders. Commissioning of on-premise/cloud resources. Well-established benchmarks to test and validate chosen hardware, possibly extending existing benchmark suites such as HEPScore. Complementary procurement/commissioning iterations based on evolving project needs	Resources are in place and validated. Use cases onboarded and initial integration within IT and experiments workflows. Longer term procurement and commissioning needs understood.
24 m	Establish a common MLOps platform automating the different steps of distributed training and AutoML. Support for hybrid deployments - on-premises and public cloud - for optimal resource usage. Support and integration with industry standard tools and libraries for distributed training (i.e. PyTorch, TensorFlow, Ray, MPIOperator, …) and AutoML (i.e Katib).. Ensure adequate integration with CERN custom libraries where needed (i.e. NNLO).	First end user use cases deployed in the new platform. At least one use case deployed per technology flavor. At least one use case making use of a hybrid infrastructure.
24 m	Support for tenant (access) management, as well as reporting and showback public cloud usage per team and project. Integration with selected hyperscalers where appropriate.	Availability of multi level (team, project, group, department) reporting on resource usage.
36 m	Optimization of deployment pipelines and end-to-end workflows. Full integration with the rest of the CERN infrastructure.	Optimised algorithms are deployed and used.
48 m	Operations	The hardware and software services are in place and fully integrated within IT and experiment workflows
60 m		

<u>Task 1.2: Development framework towards fast inference of complex network architectures on LHC online systems</u>

In this task, we will work with existing expertise in the experiment collaboration on ongoing work on tools such as hls4ml, and on expertise from selected academic and industrial partners to develop ML->FPGA model synthesis tools, addressing the needs of WP2 and WP3. The work will also focus on integrating modern ML tooling while maintaining the strict latency requirements set forth by LHC experiments' online selection system. All task items are supposed to be co-developed by CERN researchers and external partners with qualified expertise on the topic.

Time	Description	Deliverable/Milestone
6 m	Demonstrator of Knowledge Distillation workflow to real-life LHC use cases	Integration in hls4ml on multiple backends
12 m	- Deployment of transformers on FPGAs - Demonstrator of Knowledge Distillation workflow to real-life LHC use cases	- Integration in hls4ml on multiple backends - Journal publication on Knowledge Distillation on Transformer use case
18 m	Support for generic Graph Neural Networks	- Improved code-generation infrastructure to support general graphs on multiple hls4ml backends - Journal publication on Graph NN fast inference
24 m	- Support for generic Transformer network - Mid-point hls4ml release	- Journal publication describing novel hls4ml functionalities and example applications - Tutorial describing new hls4ml functionalities
36 m	ASIC-oriented development and support for novel AI-specific hardware	- Prototype on specific AI hardware (to be identified) - Journal publication - Hosting the FastML workshop at CERN
48 m	Extended integration of common operators for AI engine	- Demonstrator on specific AI hardware (to be identified) - Journal publication
60 m	- Multi-FPGA support, for inference and optionally for training - Final hls4ml release	- Tutorial describing new hls4ml functionalities - Final hls4ml release - Journal publication

<u>Task 1.3: Hardware-aware AI optimization</u>

This task will focus on leveraging external expertise on network architecture search (NAS) from selected academic and industrial partners to develop the software infrastructure needed to enable hardware-aware neural network training workflows. This work will enable the development and deployment of hardware-optimal AI-based real-time algorithms at CERN, as described in WP2 and WP3. All task items are supposed to be co-developed by CERN researchers and external partners with qualified expertise on the topic.

Time	**Description**	**Deliverable/Milestone**
6 m	Baseline development: large-scale training and optimization workflow on at least one end-to-end training library (Pytorch/Tensorflow)	Integration of the developed algorithms on the NNLO library (large-scale training package for CERN custom training workflow on HPC infrastructure)
12 m	Support of optimal workflows for hardware-aware pruning techniques with resource estimation.	- Demonstrator of network training and architecture scan for a concrete benchmark use case from WP2 or WP3 - NNLO tutorial showcasing novel functionalities - Journal publication
18 m	Support for Knowledge Distillation at training	integration of the developed compression workflows in the NNLO library
24 m	- AutoML-like flow towards automatic optimization of quantization and pruning at training time - Application of hardware-aware training on real-life use cases from WP2 and WP3	- Mid-point NNLO software release - Journal publication - NNLO tutorial showcasing novel functionalities
36 m	Hardware-aware NAS with quantization and sparsity	- Journal publication - NNLO tutorial showcasing novel functionalities
48 m	Extension of AutoML-like flow towards hardware-consumption prediction at training time	- Journal publication - NNLO tutorial showcasing novel functionalities
60 m	- Consolidation of ecosystem of compression models for edge deployments - Application of hardware-aware training on real-life use cases from WP2 and WP3	- Final NNLO release - Demonstrator of real-life use case from WP2 and WP3 - Journal publication

Task 1.4: Tensor Networks for Quantum Systems.

This task will develop and apply quantum-inspired methodology, in particular Tensor Network algorithms, to simulate quantum many-body problems unreachable by classic approaches and benchmark future applications of quantum hardware on low-entangled systems to O(100) qubits, progressing towards the development of a software stack for quantum machine learning model design, simulation, and deployment.

Time	**Description**	**Deliverable/Milestone**
12 m	Benchmark execution performance of quantum circuit simulations using tensor networks and state-vector techniques.	Setup quantum simulation techniques on classical hardware for the IT infrastructure described in 1.2 Develop a tensor network algorithm for quantum simulations with more than 37 qubits. **External** contribution: setup and benchmarking of TN/QC simulators
24 m	Compare the performance of tensor networks and state-vector simulations on large qubit systems.	Develop and deploy quantum circuit simulations for large systems, up to O(100) qubits, using tensor networks and state-vector techniques. **External** contribution: setup and benchmarking of TN/QC simulators
36 m	Benchmark the performance of quantum machine learning models to classical AI counterpart.	Develop quantum machine learning models for trigger applications with the possibility of deployment on large systems.

48 m	Achieve efficient and reliable quantum machine learning models for supervised and unsupervised learning.	Publish a study of the quantum machine learning foundations: from model building to efficient optimizers.
60 m	Determine to which extent FPGAs are beneficial for quantum circuit simulation in comparison to CPUs and GPUs.	Prototype of tensor network and state-vector simulators on FPGA using hls4ml.

Task 1.5: New computing strategies for data modeling and interpretation

From a computing point of view, the technical aspects of this work package are aligned with what is described in WP2 and WP3 in terms of code modernization on parallel architectures and with AI. The work package goals include: the porting and optimization of current event-generation codes and higher-order perturbative calculations to state-of-the-art and future hardware architectures, particularly GPUs; the development of ML/AI strategies to accelerate and improve the efficiency of phase-space sampling and the estimation of matrix elements driving the events' unweighting; AI-driven modeling of the non-perturbative aspects of proton collisions, such as the underlying event, the hadronization and the multi-parameter tuning of shower-evolution algorithms; the development of software and algorithms for efficiently exploiting next-generation computer architectures (e.g., NVidia Grace Hopper) for use in LQFT simulations on extreme-scaling low-latency/high-bandwidth accelerator-based clusters.

Advanced HPC tools, in addition to the Neural Networks already exploited, will also be needed to accelerate ancillary tasks such as the global fitting of parton density functions (PDFs), which require CPU-expensive higher-order calculations. Further work, of direct impact on the trigger studies, includes the optimization of theoretically robust clustering algorithms portable to FPGAs.

Time	Description	Deliverable/Milestone
12 m	Finalize the definition of goals and the overall coordination/sequence of the task implementation - Provide physics validation and benchmarking support for MC codes - Provide benchmarking support with lattice QFT codes guiding hardware procurement and commissioning for HPC hardware.	Organization of several community Workshops Work with Task 1.1 on devising most stringent and demanding benchmarking software. Develop LQFT benchmarking software tailored to hardware infrastructure procured under 1.1. Share expertise on parallelism and accelerator-based algorithms with TH/IT/CMS/ATLAS Define benchmarking framework for MC generators and higher-order calculations Optimisation of software algorithms for LQFT, MC generators and higher-order calculations, on latest generation classical GPU and CPU hardware
24 m	- Evolve current event generation strategies and codes, to improve performance via porting and optimization to new hardware architectures, and exploiting ML/AI strategies - Devise and execute hardware commissioning upon delivery & installation. - Port code and optimize LQFT simulation performance for new hardware that is being procured.	- Acceleration of matrix element evaluation - sign-off of new HPC cluster assuming new hardware performing to specs - show optimised parallel scaling performance of LQFT codes on new hardware Optimisation of software algorithms for LQFT, MC generators and higher-order calculations, on latest generation classical GPU and CPU hardware
36 m	- Evolve current event generation strategies and codes, to improve performance via porting and optimization to new hardware architectures, and exploiting ML/AI strategies. - Optimise LQFT simulation algorithms for procured and future hardware. - Improve LQFT simulation algorithm performance in large-volume and investigate variance-reduction techniques for hadronic matrix elements.	- Acceleration of phase space generation. - Show LQFT code-performance improvement. - Improvements of signal-to-noise ratio in hadronic matrix elements Optimisation of software algorithms for LQFT, MC generators and higher-order calculations, on latest generation classical GPU and CPU hardware

48 m	- Evolve current event generation strategies and codes, to improve performance via porting and optimization to new hardware architectures, and exploiting ML/AI strategies - Develop variance-reduction techniques for hadronic matrix elements in LQFT in large volume and combine with spectral-reconstruction techniques.	- Acceleration of parton shower evaluation - Improvements of signal-to-noise ratio in hadronic matrix element and study of systematics in spectral reconstruction Optimisation of software algorithms for LQFT, MC generators and higher-order calculations, on latest generation classical GPU and CPU hardware
60 m	- Validation of code improvements -Extension of code improvements to the global ecosystem of event modeling tools -Develop understanding of viability of spectral reconstruction techniques.	- Event mass production and benchmarking - Applications to PDF fitting, MC parameter tuning, evolution of jet clustering algorithms and FPGA implementation - application of LQFT to spectral-reconstruction problems.

<u>Task 1.6: New Physics scenarios and Standard Model properties as trigger benchmarks</u>

To evaluate the impact of the next-generation triggers on the enhanced sensitivity to New Physics scenarios and on the determination of fundamental properties of the Standard Model (SM), theorists will develop benchmarks in close collaboration with the experiments. These will include concrete models that extend the SM as well as anomaly searches that are as model-independent as possible to reduce the theory bias of model building. In addition, theorists will determine to which extent the next-generation triggers improve the precision of the determination of SM parameters.

These concrete physics targets will allow for a robust performance assessment and validation of the next-generation triggers. Furthermore, they will serve as a guidance for the trigger optimization for relevant use cases. As a spin-off of the required simulation activities, algorithmic improvements of event-generation codes, including porting to new HPC architectures, will be explored.

The activity will start during the second year of the and last for 3 years.

Time	Description	Deliverable/Milestone
12-24 m	- Defining clear NP observables of specific relevance to new triggers, with a clear understanding of performance of current triggers	- Development of New Physics benchmark scenarios with exotic detector signatures - Evaluation of the current sensitivity based on the current triggers
36 m	Liaise across the collaboration to determine planned next generation trigger performance.	Development of SM benchmarks Cooperation with experimentalists from ATLAS and CMS on the next-generation triggers and evaluation of the improved performance on the benchmarks
48 m	Benchmark new trigger performance.	Model-independent anomaly search and assessment of the new trigger performance Comparison of current and next-generation triggers on precision SM measurements (eg Higgs boson couplings)

Task 1.7: Common software developments for heterogeneous architectures

To make efficient use of accelerator (GPU and FPGA) devices in the software designed for the High-Luminosity LHC, various common developments and improvements are needed in the frameworks and code bases of the experiments and Monte Carlo generators. Frameworks need to make efficient use of all available computing resources of single compute nodes, and even possibly multiple nodes at the same time. Existing implementations should be harmonized between the experiments, and optimization efforts should be shared.

In order to make the software used by the experiments, TH and IT as efficient as possible, we need to develop techniques and improvements in the following areas:

- Efficient scheduling of computing steps on heterogeneous devices for ensuring that CPU and accelerator resources are maximally utilized;
- Efficient data structures for heterogeneous software to ensure that memory copies are minimized and are as efficient as possible;
- Common High-Energy Physics libraries for heterogeneous systems for implementing optimal calculation of common operations for HEP code, running on accelerator devices;

- Efficient interfaces to Machine Learning inference engines to minimize data movements and execution latencies;
- Finally, novel programming languages should be evaluated for the implementation of reconstruction algorithms in the High Level Triggers.

Time	**Description**	**Deliverable/Milestone**
12 m	Develop realistic demonstrators, tests and examples for all of the projects. Implement “standalone” workflows with ≥2 use cases for the framework scheduling. Put the first EDM prototype in place. Put test cases for the small vector/matrix operations in place. Investigate all target programming languages for their interoperability with C++.	Produce a report about the status of the preparatory steps of the different projects.
24 m	Develop more backends, tests and use cases in all of the projects. The scheduling workflows are fully implemented using their “native” scheduling techniques. Benchmarking is put in place for the small vector/matrix test code. Benchmarking is implemented for the first EDM prototype. Investigate the ease of accelerator use in all studied programming languages.	Produce a report about the status of the demonstrators and the results achieved by that point.
36 m	Start with the ML inference work and complete all of the demonstrators and prototypes. Have all scheduling workflows implemented using all of the studied scheduling techniques. Small vector/matrix linear algebra interfaces and at least one backend are put in place. Agrees with industry partner(s) on new/optimized ML inference interface(s). Port one of the scheduler prototypes to the most promising novel programming language.	Produce a report about the status of developments and about the negotiations with 3rd parties.
48 m	Complete benchmarking on all projects and perform the necessary optimisations arising from the benchmark results. All scheduling techniques are benchmarked with all of the workflows. Small vector/matrix algebra library benchmarked on all tests with all backends. New ML interface(s) is/are benchmarked and optimized on a realistic	Produce a report about the performance results of all of the projects.

Time	Description	Deliverable/Milestone
	workflow. The EDM implementation is fully benchmarked and its interoperability with I/O and reflection is tested. Investigate further programming languages.	
60 m	Write up all of the findings of all of the projects in multiple conference proceedings / documents. Help experiments with implementing new code / techniques in their software.	Produced multiple write-ups that document all of the R&D.

Work Package 2: Enhancing the ATLAS Trigger and Data Acquisition

This work package focuses on an enhancement of the already ambitious upgrade of the ATLAS experiment's trigger and data acquisition system for the High Luminosity Phase of the LHC (HL-LHC) scheduled to start in 2029. Novel approaches to trigger event selection will be developed that will extend the ATLAS physics potential, in particular exploiting state-of-the-art Machine Learning techniques. Efficient use of acceleration technologies will be investigated to extend the capabilities of the trigger and data acquisition system, while improving the system's energy efficiency. A successful implementation of this work package will result in the collection of richer collision events, extend the experiment's sensitivity to a broader range of new physics beyond the Standard Model scenarios, and enable ATLAS to exploit state-of-the-art processing architectures in its online event selection to achieve the best possible physics performance at HL-LHC.

Work Package Management:

Time	Description	Deliverable/Milestone
Total over 5 years (rounded)	WP coordination and technical supervision; coordination with project management and external contributors	Technical reports on achieved deliverables and milestones

Task 2.1: Optimal Real-Time Event Selection in the Global Trigger system

The Level-0 Global Trigger (L0Global) is a new subsystem, which will execute offline-like reconstruction algorithms on full-granularity calorimeter data in real time at 40 MHz, with latency of a few microseconds and data throughput of 50 Tbps. Novel Machine Learning based reconstruction and feature extraction algorithms need to be developed to extend the physics potential of the experiment.

We will develop a common framework for optimizing such algorithms in four main areas of the Global Trigger: electrons/photons, taus, jets, and multi-object, full-event reconstruction; as well as for the preprocessing of the calorimeter data inputs. The ultimate goal is to develop and optimize these algorithms all the way to full firmware implementation and integration into the Global Trigger system to augment or replace the baseline algorithms during Run 4 of the LHC.

Time	**Description**	**Deliverable/Milestone**
12 m	Develop framework/toolkit for optimizing ML algorithms in terms of physics performance vs. FPGA resource vs. latency.	Groundwork for the development and optimization work.
24 m	Design initial variants of ML algorithms with limited input features. Design and deployment of L0Global testbed for the development, testing and integration of the ML algorithms, in parallel with the baseline project, and for the training of the fellows and students of the team.	Testbed available
36 m	Implement and optimize algorithms for the global trigger environment. Integrate with global trigger framework with modified data processing path. Develop trigger simulation of initial algorithm variant.	Initial implementation as standalone algorithms in development boards.
48 m	Extend algorithm design to consider extended feature space. Implement/optimize new algorithm variant(s). Extend trigger simulation to new algorithm variant(s) & study and possibly extend physics performance.	Algorithms ready for full framework implementation. Demonstrate the added value of this project.
60m	Integrate optimized algorithm variants with global trigger and demonstrate operational capability.	Ready to go live in the ATLAS Trigger System during Run 4

Task 2.2: Enhancing the Level-0 Muon Trigger

We propose to develop fast algorithms on dedicated hardware for the Level-0 (L0) muon trigger system to cover use cases that are not part of the ATLAS baseline system for the HL-LHC. In the barrel region of ATLAS, the RPC detector technology is used to provide muon trigger candidates to the L0 MDT trigger processor where those candidates are either rejected or refined by making use of the superior spatial resolution of the MDT chambers. To improve the robustness of the L0 muon trigger system against the potential loss of performance due to aging RPC detectors, we propose extending the baseline system to rely on a smaller number of RPC chambers (as little as one) to provide seed triggers. This would significantly increase the number of candidates to be confirmed by the L0 MDT trigger processor. An additional goal of this proposal is to trigger directly on non-pointing signatures from the decay of long-lived exotic new particles. To achieve those goals, new algorithms need to be developed, possibly exploiting Machine Learning techniques, to fit within the hardware resources to perform the required muon trigger tasks within a maximum latency of 1.3 micro-seconds.

Time	Description	Deliverable/Milestone
12 m	Conceptual study of different extensions of the L0 MDT trigger: RPC seed, addition of Tile seed, exotic signatures. Develop initial algorithms and estimate performance.	Understand the challenges and potential solutions for each of the scenarios.
24 m	Finalize and implement algorithms in firmware for the RPC seed options. Design and production of prototype hardware for the enhanced muon trigger	Evaluate latency and resource limitations. Prototype hardware available
36 m	Finalize and implement algorithms in firmware for exotic signatures option. Develop initial algorithms for the MDT standalone option and estimate performance.	Evaluate latency and resource limitations.
48 m	Revise and optimize algorithms for RPC + Tile seed options given the baseline hardware resources. Integration and testing on baseline hardware. Final performance evaluation.	Algorithms ready for full framework implementation.
60m	Revise and optimize algorithms for exotic signatures. Integration and testing on baseline hardware. Final performance evaluation.	Final evaluation of hardware resource needs for those options.

Task 2.3: High Throughput Data-Collection

To fully exploit the physics potential of the novel trigger approaches developed in WP2, the ATLAS data acquisition infrastructure needs to provide all the required event information reliably and within minimal latency. This so-called Readout, is a challenging data-acquisition subsystem, responsible for interfacing detector-specific optical links to a commercial network. The goal of this task is to optimize the readout performance and to address bottlenecks in the system. Extending the information content of the data provided by the detector may be required to maximize the physics performance of the Level-0 selection.

Time	**Description**	**Deliverable/Milestone**
12 m	Perform the team hiring and set the basis for the project. In particular perform a market survey of the existing technologies and understand the requirements in terms of hardware infrastructure.	Team hiring. Downselection of promising technologies. Definition of the required lab setups.
24 m	The hardware platform required for the technology evaluation have to be specified, procured and configured and installed. In parallel software development of prototypes with the selected technologies as well as benchmarking tools can start. Initial comparative measurements will be performed.	Procurement, deployment and configuration of the test-benches. Initial comparative measurements. Prototype available
36 m	Promising technologies will be retained. The associated prototypes will evolve into proof of concepts. This may require scaling up the hardware infrastructure as well as the benchmarking tools.	Scale-up of test infrastructure. PoC
48 m	Following satisfactory results with PoCs, the valuable technologies have to be ported into the baseline DAQ infrastructure that has been developing in parallel. Large scale tests, possibly on the production system as it is being assembled, will be organized.	Integration in the DAQ infrastructure. Validation on large scale (e.g. production system)
60 m	Contribute to the integration and commissioning of the DAQ system	Deployment and commissioning of selected technologies.

Task 2.4: Event Filter Tracking

The goal of this task is the development of an algorithmic solution for the ATLAS Event Filter track reconstruction, employing optimal classical numerical and Machine Learning techniques, and to deploy it on the most suitable hardware architecture. Machine Learning approaches to tracking, as Graph Neural Networks, will be investigated to replace parts of the (or possibly the full) classical numerical algorithm chain. The aim is to optimize the physics and processing performance of the track reconstruction and to investigate the potential of porting parts of the tracking chain on systems with co-processors like GPUs and FPGAs. This task will implement the tools developed in WP 1.2: Development framework towards fast inference of complex network architectures on LHC online systems and in WP 1.7: Framework integration of accelerators and provide feedback on their performance for further optimization.

Milestones for Task 2.4:

Time	Description	Deliverable/Milestone
12 m	Review of existing Event Filter Tracking approaches, with particular emphasis on advanced Machine Learning methods. Implementation of suitable Machine Learning based models and contribute to the AI/ML optimisations for Event Filter Tracking.	Identify areas of potential improvements among existing approaches and possible complementary methods.
24 m	Work towards the prototype implementation (algorithmic and hardware configurations) with throughput and performance optimisation of the system. Implement tools from WP 1.2 and 1.8.	Performance evaluation of optimized AI/ML tracking algorithm. Provide feedback to WP 1.2 and 1.8.
36 m	Contribute to the final system preparation; optimisation of system design performance. Evaluate improvements from updated tools developed in WP 1.2 and 1.8.	Partial system deplo1 FTE (Staff) 2 FTE (Grad) 2.5 PhD/PJASyment, demonstrating scaling to full EF tracking system. Provide feedback to WP 1.2 and 1.8
48 m	Optimize and finalize the implementation of the system into the ATLAS TDAQ infrastructure. Implementation of final tools developed in WP 1.2 and 1.8.	Demonstration of potential of algorithm and computational improvements; full Event Filter Tracking system deployment.
60 m	Tune algorithm performance and rejection for trigger needs. Contribute to the consolidation and commissioning of the system.	Coherent description of the project in documentation and publication.

Task 2.5: Optimized Event Filter Muon Trigger Selection

The goal of this task is to fully exploit the extended coverage of the Level-0 muon trigger (Task 2.2) and the novel tracking infrastructure (ACTS) developed in Task 2.6 to improve the physics performance of the Event Filter muon track reconstruction. Migrating to ACTS should significantly reduce the computing resources needed for muon reconstruction and potentially provide an enhanced precision for the (combined) muon track fitting. Using ACTS will also facilitate porting parts of the muon track reconstruction onto accelerator hardware. Novel pattern recognition algorithms using Machine Learning may further improve the efficiency and technical performance of the muon reconstruction. We will also study the potential of porting such novel algorithms onto accelerators like GPUs or FPGAs. As a result ATLAS will be able to handle the increased rate of Level-0 muon candidates within the available processing resources to retain more interesting events with muons in the final state.

Time	Description	Objectives
12 m	Measure the performance of the existing Muon Event Filter.	Understanding of the current algorithm bottlenecks and hotspots.
24 m	Migrate current MS EF to use ACTS (via wrapping or complete rewrite within ACTS). This will involve working on components shared with the offline (non-trigger) reconstruction. Together with Task 2.6, evaluate potential novel AI reconstruction techniques.	First prototype of ACTS/AI Muon Event Filter running on simulated datasets.
36 m	Performance evaluation of prototype new reconstruction algorithms. The algorithms will be compared to the existing algorithms in terms of efficiency, misidentification rates and throughput.	Compare performance of prototype new reconstruction algorithms. Identify which algorithms should be further developed. Input to ATLAS decision on final hardware design.
48 m	R&D on selected algorithms to improve performance (efficiency, misidentification rates and throughput). Development of selected AI algorithms (together with Task 2.6) on selected hardware platforms. Integration into ATLAS HLT infrastructure.	Updated AI reconstruction running in ATLAS trigger. Algorithm robustness & documentation.
60 m	Deployment and commissioning of the system.	Final algorithm performance estimates. Algorithms running in ATLAS partition and being exercised on cosmic data.

Task 2.6: Common Tracking Event Filter infrastructure

In this task, the infrastructure for the integration of the novel Event Filter track reconstruction into the ATLAS software ecosystem is developed. To facilitate exploiting heterogeneous processing architectures for the Event Filter track reconstruction, the infrastructure needs to enable efficient offloading of track reconstruction algorithms onto accelerator hardware. Within the ATLAS software ecosystem the Event Filter tracking is foreseen to run embedded in the common tracking software ACTS, an open-source component library for charged particle reconstruction shared amongst several experiments. This task will enable the integration of newly developed tracking algorithms and software modules (including the enhanced use of machine learning based solutions) as indicated in Tasks 2.4 and 2.5.

Time	Description	Deliverable/Milestone
12 m	Review status of the ACTS CPU baseline and outcome of the ACTS R&D line on parallelization, identify and define	Arrive at an understanding of work areas that have no, partial or full solution coverage. Definition of candidate algorithmic pipelines and necessary

Time	Description	Deliverable/Milestone
	showcase examples for data transfer pattern and algorithm execution and optimize them. Execute first test examples with different host/device workloads.	support software.
24 m	Establish EF Track reconstruction pipelines in various flavors with no, full and partial offloading, including accelerator-friendly modules, such as the GNN based track finding candidate within the ACTS umbrella.	Proof of principle showcase examples with different host/device workloads. Identification of non-beneficial R&D directions.
36 m	Development of an Integration strategy of an enhanced ACTS software suite with adapted event data model and execution scheduling into the ATLAS software framework, while ongoing algorithmic improvements.	Runnable examples using ACTS infrastructure within the ATLAS software ecosystem.
48 m	Contribute to the integration of support software and ACTS based track reconstruction developed within this work package into the ATLAS software framework, dedicated adaption to Event Filter application.	Contribution to an enhanced ATLAS software system with increased heterogeneous computing support.
60 m	Documentation, performance evaluation and further optimisation of the developed software solutions for enhanced heterogeneous computing within ACTS and ATLAS.	Final project report and publication describing the implemented strategies, the developed software and quantifying their eventual impact.

Task 2.7: Enhanced Reconstruction for Higher Level Event Filtering

Scenarios of physics beyond the Standard Model, as those developed in WP1, can produce signals in the experiment which require novel triggering, data acquisition and analysis techniques to be detected. The goal of this task is to exploit the enhanced functionality and performance of the novel tracking approaches, including those developed in Tasks 2.4 and 2.5, to extend the physics potential of ATLAS and to develop novel algorithmic approaches to efficiently search for non-standard particle signatures. Further extensions to improve reconstruction algorithms combining tracking and other detector information using innovative algorithmic approaches will be investigated to improve the sensitivity of the Event Filter selection. Dynamic bandwidth allocation and algorithm parameters to account for unforeseen physics signals and changes to detector and accelerator running conditions using advanced machine learning techniques will also be explored. With this task we aim to benefit from the theoretical studies done in Task 1.6.

Time	Description	Deliverable/Milestone
12 m	Identify promising use-cases to benefit from enhanced EF reconstruction, by evaluating existing physics studies and performing new evaluations of interesting scenarios.	Identify where enhanced EF reconstruction can yield largest benefits to physics sensitivity.
24 m	Perform initial characterisation of potential gains to select a sub-set of signatures for a full implementation, by emulating "ideal" trigger performance in simulated physics analyses. Establish framework for dynamic bandwidth allocation	Understand achievable improvements. Converge on list of signatures for further implementation. Obtain basis for future extension of bandwidth management.
36 m	Implementation of prototype trigger signatures, assuming offline-like reconstruction capabilities. Refine estimate of physics performance gains based on prototype implementation. Improve algorithms based on physics sensitivity. Assess dynamic algorithm tuning in the process.	Demonstrate feasibility of algorithms. Develop optimal approaches for physics use cases.
48 m	Full implementation of trigger signatures within the scope of the enhanced EF reconstruction. Exploit developments achieved in other tasks. Study support of dynamic tuning for existing triggers.	Obtain functional EF reconstruction algorithms with adequate computational performance
60 m	Final validation and deployment of trigger signatures, potential dynamic menu and algorithm features. Fine-tuning of computational performance.	Finalize algorithms and deploy to the HL-LHC trigger menu. Ensure compatibility with the wider trigger menu.

Work Package 3: Rethinking the CMS Real Time Data Processing

This working package aims to rethink the CMS data acquisition system allowing the CMS physics program to operate over all the collisions produced by the LHC. This is achieved through the High-Level Trigger (HLT) Real-time Reconstruction Revolution (R^3) and a novel L1-trigger scouting stream. These goals are achieved by mixing traditional physics-inspired algorithms with cutting-edge AI solutions and leveraging synergy between CMS physicists and data scientists. Proof-of-concept R&D projects on these activities are ongoing. Some of them already established the validity of the ideas behind the proposed tasks.

The tasks listed below are independent efforts that will be carried on in parallel across a four-year period. Tasks 3.1, 3.3, and 3.4 will deliver intermediate milestones on a two-year time scale that could be deployed in the CMS HLT during Run3. Similarly, part of the work in 3.5, 3.6, and 3.7 will focus on developing the L1 trigger for Run3, with the possibility of deploying next-generation algorithms on a two-year time scale.

Specific material for CMS HLT tasks 3.1-3.4

This is specific material, covering needs for all CMS HLT tasks 3.1.1-3.4

- [1st year] 50k for an R&D machine with different GPUs
- [3rd year] 50k for an R&D machine with different GPUs

Task 3.1.1: The Real-time Reconstruction Revolution (R^3 - Rcube)

The traditional CMS Phase-2 reconstruction is off-line, takes tens of second per event, and is based on algorithms developed decades ago. The R^3 project will aim to modernize this system by leveraging capacity across the data center, heterogeneous compute resources, and modern AI-driven techniques, using modern development methodologies, allowing for more accurate event reconstruction and higher confidence in trigger decisions.

Time	**Description** **Deliverable/Milestone**
6 m	● Written report outlining the bottlenecks in the existing **particle flow and jets/MET** reconstruction software, and proposing solutions to improve its performance. Create a public website to showcase the project, featuring a general project description and a visually appealing logo. Additionally, establish a dedicated section within the website to provide more detailed information regarding **jet/MET** and **particle flow**. ● Written report outlining the bottlenecks in the existing **stand-alone and global muon** reconstruction software, and proposing solutions to improve its performance. Establish a dedicated section within the website to provide more detailed information regarding **muon** reconstruction. ● Written report outlining the bottlenecks in the existing **electromagnetic objects** reconstruction software, and proposing solutions to improve its performance. Establish a dedicated section within the website to provide more detailed information regarding **electromagnetic** objects reconstruction. ● Written report outlining the bottlenecks in the existing **tau leptons** reconstruction software, and proposing solutions to improve its performance.
12 m	● Have a comprehensive validation suite to assess the physics and computational performance of the **particle flow and jets/MET** reconstruction algorithms. The suite must cover the heterogeneous architectures of interest to CMS. ● Have a comprehensive validation suite to assess the physics and computational performance of the **stand-alone and global muon** reconstruction algorithms. The suite must cover the heterogeneous architectures of interest to CMS.

	● Have a comprehensive validation suite to assess the physics and computational performance of the **electromagnetic objects** reconstruction algorithms. The suite must cover the heterogeneous architectures of interest to CMS. ● Have a comprehensive validation suite to assess the physics and computational performance of the **tau leptons** reconstruction algorithms. The suite must cover the heterogeneous architectures of interest to CMS.
30 m	● Experimental prototypes of the data-structures and algorithms for the **particle flow and jets/MET** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software. These algorithms will be designed for optimal performance in heterogeneous computing. ● Experimental prototypes of the data-structures and algorithms for the **stand-alone and global muon** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software. These algorithms will be designed for optimal performance in heterogeneous computing. ● Experimental prototypes of the data-structures and algorithms for the **electromagnetic objects** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software. These algorithms will be designed for optimal performance in heterogeneous computing. ● Experimental prototypes of the data-structures and algorithms for the **tau leptons** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software. These algorithms will be designed for optimal performance in heterogeneous computing.
36 m	● Validation of the physics and measurement of the computational performance of the **particle flow and jets/MET** experimental reconstruction algorithms on diverse heterogeneous computing platforms. ● Validation of the physics performance of the **stand-alone and global muon** experimental reconstruction algorithms on diverse heterogeneous computing platforms. ● Measurement of the computational performance of the **stand-alone and global muon** experimental reconstruction algorithms on diverse heterogeneous computing platforms. ● Validation of the physics performance of the **electromagnetic objects** experimental reconstruction algorithms on diverse heterogeneous computing platforms. ● Measurement of the computational performance of the **electromagnetic objects** experimental reconstruction algorithms on diverse heterogeneous computing platforms. ● Validation of the physics performance of the **tau leptons** experimental reconstruction algorithms on diverse heterogeneous computing platforms. ● Measurement of the computational performance of the **tau leptons** experimental reconstruction algorithms on diverse heterogeneous computing platforms.
54 m	● Production-grade data-structures and algorithms for the **particle flow and jets/MET** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software. ● Production-grade data-structures and algorithms for the **stand-alone and global muon** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software.

	● Production-grade data-structures and algorithms for the **electromagnetic objects** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software. ● Production-grade data-structures and algorithms for the **tau leptons** reconstruction exploiting a data-oriented approach, integrated in the CMS reconstruction software.
60 m	● Deployment and commissioning of the whole infrastructure, in preparation for the HL-LHC data-taking.

Task 3.1.2: R³ optimized data structures for heterogeneous platforms

The development of data-oriented structures ("Structure of Arrays", SoA) will be fundamental for R³ to reach its goal. This data representation can achieve better memory bandwidth and vectorization performance for classical algorithms and provide a seamless interface to AI algorithms. Its adoption in the HEP software stack requires the development of a user-friendly, generic SoA implementation. To achieve the best performance running real-time trigger selection, the I/O subsystem of the CMS framework will be extended to leverage direct data transfers between the network and storage subsystems on one side, and the accelerators on the other, bypassing the host CPU.

Time	**Description/Deliverable/Milestone**
12 m	● Collection of data formats use cases from the CMS software production version, classification thereof, identification of the necessary features of the generic CMS SoA, collection of feedback from stakeholders and proposal of an implementation plan.
24 m	● Implementation of the SoA, replacement of the initial set of data formats and integration in CMSSW based on the advancement of the evolution of algorithms re-invented for accelerators. Investigation of existing technologies for direct I/O to-from device in a portable fashion, with focus on ROOT.
36 m	● Conversion of more data formats to SoAs and integration in CMSSW, proposal of a design upgrade of classical and RNtuple ROOT I/O to accommodate persistification of data present on the device to storage and efficient recreation of SoAs in device memory starting from storage.
48 m	● Implementation of the features necessary for direct ROOT I/O from/to accelerators and Development of a benchmarking suite for the new I/O based on real life use cases of CMS.

6 0m	● Optimization of the new functionalities to obtain the highest throughput, considering reading from local storage, mass storage pools or remote reads through the xrootd technology.

Task 3.2: Evolving the CMS experiment software into a client-service distributed application for HLT

In this task, the CMS data processing framework is extended to become capable of adapting to different network topologies to leverage remote accelerators, with little or no modification to the core code.

Time	**Description/Deliverable/Milestone**
12 m	● Implementation of a client-server, multithreaded, distributed test application, based on the CMS software framework, leveraging high-speed host-to-host or shared memory communication.
18 m	● Implementation of a small-scale demonstrator of a full HLT-like application.
30 m	● Support for optimal use of remote accelerators, e.g. using RDMA to/from GPU memory, is integrated in the CMS software framework.
36 m	● Support for multiple servers and distributed configurations is integrated in the CMS software framework.
48 m	● Delivery of a report comparing the approaches developed in the CMS software framework to improve the resiliency of the system, such as server redundancy and client-side failure mitigation strategies.
54 m	● Delivery of a report evaluating the performance of different network interconnects and communication protocols.
60 m	● Large scale deployment and testing of the whole infrastructure in view of the readiness for the 2029 HL-LHC data-taking.

Task 3.3: Reduction of the RAW data size for HLT

In this task, multiple approaches to the compression of RAW data are characterized, with different trade-offs between the compression factor, latency, available hardware/detector infrastructure and impact on the final physics result. The goal would be achieved by considering both lossy and lossless algorithms, as well as replacing basic information with higher-level quantities derived from it (physics-driven compression).

Time	Description/Deliverable/Milestone
6 m	● Production of a report illustrating the impact in terms of RAW data size coming from each detector, and suggesting the major area of interest/intervention.
18 m	● Assessment of two different approaches to data compression: lossless compression on accelerators, and replacement of part of the RAW data with low-level reconstructed quantities (RAW').
30 m	● Implementation of the most promising solutions replacing part of the RAW data with low-level reconstructed quantities in the CMS reconstruction software.
42 m	● Assessment of the trade-off between the data reduction and the impact on the high-level physics reconstruction, as part of the RAW data are replaced by high-level reconstruction objects.
54 m	● Implementation of the most promising solutions replacing part of the RAW data with high-level reconstructed quantities in the CMS reconstruction software.
60 m	● Large scale deployment and testing of the whole infrastructure in view of the readiness for the 2029 HL-LHC data-taking.

Task 3.4: Optimal calibration for HLT

This task will optimize the calibration process for the CMS detectors, from hours currently, to an on-line predictive model leveraging AI techniques. Such an improvement is essential to push real-time analysis based on R3 software (task 3.1) to the same accuracy that we typically achieve off-line. One could then store high-quality high-level information, which implies a big save in terms of storage.

Time	Description/Deliverable/Milestone

12 m	● Production of a report illustrating the current calibration workflows in CMS and evaluating the impact of non-optimal calibrations on the physics performance of the online reconstruction.
24 m	● Creation of a small-scale prototype that would buffer 30% of the Run-3 RAW data for about 8 hours, derive improved calibrations, and use them for the online reconstruction of the “HLT scouting” data stream.
48 m	● Production-grade calibration system is integrated in the CMS online infrastructure and tested in a realistic environment. Production of a report assessing the impact of improved online calibrations on the physics performance of the HLT and of the low-level reconstruction objects stored in the RAW’ format (see Task 3.3)
60 m	● Large scale deployment and testing of the whole infrastructure in view of the readiness for the 2029 HL-LHC data-taking

Task 3.5: L1 scouting for HL-LHC

We propose to develop the hardware architecture and the algorithms of a real-time analysis facility that would be operated at 40 MHz during HL-LHC. Leveraging external expertise on task-oriented optimization of heterogeneous computing environments, this task will operate a survey of hardware solutions to support a complete portfolio of physics analyses.

To demonstrate real time analysis at 40MHz and benchmarks possible solutions, a test stand will be set up with FPGAs, servers hosting different accelerators and high-speed optical networking, connected to prototype CMS L1 trigger boards. As part of the project, a new ATCA data acquisition board for L1 scouting will also be developed, with higher input bandwidth and exploiting on the latest technologies (e.g. Xilinx Versal HBM, 400 GbE, …). This test stand will be used also for some of the R&D in Tasks 3.6 and 3.7.

Time	**Description**	**Deliverable/Milestone**
6 m	● Start development of prototype physics analyses with students ● Define and procure hardware infrastructure for scouting demonstration systems (servers, networking equipment, …) and FPGA development kits ● Survey accelerator technologies (GPU, GPU+NIC, FPGA, ACAP)	● Analysis projects defined, students recruited ● Procurement of hardware infrastructure launched ● Short list of accelerator devices to procure for R&D

12 m	● Development of prototype physics analyses with students ● Installation of hardware infrastructure for scouting demonstration systems ● Procurement of accelerators ● R&D on implementing HEP algorithms on accelerators	● Complete development of first classical analyses in simple final states: strategy, physics reach, bandwidth & processing requirements (report / conference talk) ● Hardware infrastructure installed & commissioned ● Accelerators for R&D procured ● First comparison of HEP algorithms on accelerators (report / conference talk)
18 m	● Development of prototype physics analyses with students ● Set up of complete a first small-scale demonstrator of the scouting data acquisition system with bandwidth of order 100 Gbps ● R&D on using FPGA accelerators with QSFPs and/or converged GPU+NIC cards ● Conceptual design for new data acquisition board	● First scouting DAQ demonstrator completed
24 m	● Development of prototype physics analyses and their implementation with students ● Extend scouting demonstrator adding live data processing capability ● R&D on using FPGA accelerators with QSFPs and/or converged GPU+NIC cards ● Design of new data acquisition board (components, PCB layout, …)	● Complete development of first AI-assisted analyses and analyses in complex final states ● Demonstration of running first classical analyses in simple final states on scouting system ● Completed R&D on using FPGA accelerators with QSFPs and/or converged GPU+NIC cards, produce a report, and working prototype if approach found to be viable ● Conceptual design for new data acquisition board completed ● Defined list of electronics components for new data acquisition board, and procurement started
36 m	● Development of prototype physics analyses and their implementation with students ● Extend scouting demonstrator with offloading to accelerators ● Design of new data acquisition board (components, PCB layout, …) ● Start R&D on next-generation accelerator devices	● Complete the development of the first analyses in high-rate final states, e.g. analyzing all reconstructed L1 tracks ● Demonstration of running AI-assisted analyses and analyses in complex final states ● Procurement of a few more computing nodes and next generation accelerators ● PCB design for new data acquisition board completed and sent for manufacturing, electronics components available
48 m	● Development of prototype physics analyses and their implementation with students	● Improved version some of the prototype analyses with more advanced strategy, evaluation of the physics gains and increased processing requirements

	● Assembly and testing of first prototype of next generation ATCA data acquisition board. ● Extend scouting demonstrator bandwidth & throughput ● Continue R&D on next-generation networking solutions & accelerator devices	● Scouting demonstrator extended to a bandwidth of order 400 Gbps ● First prototype of new data acquisition board assembled and tested. ● Complete R&D on next-generation networking solutions, and define solution to use for final demonstrator system ● Complete R&D on next-generation accelerator devices, and define solution to use for final demonstrator system ● Procurement of remaining components to finalize demonstrator
60 m	● Implementation of prototype physics analyses ● Finalization of design for new data acquisition board, possible new iteration of production and assembly ● Extend scouting demonstrator bandwidth & throughput, using findings from previous R&D	● Scouting demonstrator extended to an input bandwidth > 1Tbps, e.g. capable of receiving all reconstructed L1 tracks ● Demonstration of running analyses in high-rate final states in the scouting system ● Demonstration of running the improved analyses from previous milestone ● Final design of new data acquisition board completedDefinition of solutions to use for final Run4 system

Task 3.6 Practical real-time AI for Level 1 Trigger and L1 Scouting

For this task, we propose to research and develop methods to make optimal use of the information that is available in the trigger system, and a system to deploy models with robust provenance tracking and reproducibility. We anticipate that Machine Learning will be prevalent throughout the CMS L1T during the HL-LHC era, with around 20 models and 50 billion inferences per second already accounted for. Developing and operating the experiment with this large amount of ML in the data acquisition pipeline is a new frontier for CMS.

Time	**Description**	**Deliverable/Milestone**
6 m	● Develop AI in Correlator for better object reconstruction and linking (electrons, photons, vertexing, jets): define problem, start work on firmware and emulator ● Start development of MLOps practices for Run 3 Global Trigger NNs	● First physics performance results for the AI algorithms in correlator

12 m	● Develop AI in Correlator for better object reconstruction and linking (electrons, photons, vertexing, jets): firmware and emulators ● Finish develop MLOps practices for Run 3 Global Trigger NNs	● MLOps practices for Run 3 Global Trigger NNs defined and documented ● Trained AI models integrated in Correlator trigger firmware and emulator
18 m	● Develop AI triggers (Global Trigger) using improved correlator objects (better efficiency at low pT without increasing rate for electron/photon trigger paths, jet tagging based triggers) ● Further develop AI in Correlator for better linking (for example electrons/photons), vertexing ● Deploy and use MLOps development from Year 1	● Physics performance results for updated AI algorithms for Correlator
24 m	● Develop AI triggers (Global Trigger) using improved correlator objects ● Further develop AI in Correlator for better linking (for example electrons/photons), vertexing ● Deploy and use MLOps development from Year 1	● AI triggers integrated in Global Trigger firmware and emulators ● Updated AI algorithms for Correlator: firmware and emulators released ● Report on experience from deployment of MLOps practices for Run 3 Global Trigger NN
36 m	● Scale up Year 1-2 MLOps development for Phase-2 ● Co-train AI between Correlator and Global Trigger (and potentially upstream e.g. HGCal, Tracker) for optimal usage and transport of information - optimal AI compression for final trigger selection	● Physics performance results for co-trained Correlator+GT AI algorithm and estimation of firmware resources ● Draft of MLOps practices for Phase-2
48 m	● Deliver and integrate into Correlator+GT firmware for Year 3 co-trained NNs, develop emulator ● Develop full scale Phase-2 MLOps demonstrator	● Firmware and emulators for co-trained Correlator+GT AI algorithm released ● Full scale Phase-2 MLOps demonstrator developed
60 m	● Integrate and commision co-trained Correlator+GT firmware in final system	● Correlator+GT firmware commissioned and deployed in final system ● Define final MLOps practices for Run 4

Task 3.7: L1 scouting data compression for efficient data acquisition and anomaly detection

L1 scouting provides the possibility for unbiased HL-LHC data acquisition and storage for future analysis, but the resulting datasets would be prohibitively large, order 100-1000 PB per data-taking year depending on the kind of information saved. In this task, we propose to apply cutting-edge compression techniques, including nonlinear lossy compression with AI algorithms (e.g., autoencoders) to reduce the L1-scouting dataset size. Autoencoders are also a promising algorithm for anomaly detection, and so they will also be explored for that purpose, that can be already applicable to Run 3 data. For this task, we would also work on optimizing the hardware design of the algorithm (resource consumption and latency), to potentially run it as part of the main L1 trigger system to add scouting, both for Run3 and HL-LHC. Collaboration with external partners with expertise with cutting-edge AI algorithms would be extremely functional to this goal.

Time	**Description**	**Deliverable/Milestone**
6 m	● Develop demonstrator of unsupervised anomaly detection with Run 3 physics collisions	
12 m	● Complete demonstrator of unsupervised anomaly detection with Run 3 physics collisions ● Optimize autoencoder-based off-the-shelf model to increase out-of-distribution detection robustness (e.g., via custom losses or self-supervised domain adaptation)	● Demonstrator of unsupervised anomaly detection with Run 3 physics collisions delivered
18 m	● Develop end-to-end data analysis strategy that include the high-level trigger and offline statistical analysis steps ● Develop unsupervised graph-based ML model optimized to meet inference latency and resource constraints of the high-level trigger system ● Develop method for offline characterization of anomalies to deliver physics results (e.g., through clustering algorithms or semi-supervised strategies)	
24 m	● Develop end-to-end Run 3 data analysis strategy for that include the	● End-to-end Run 3 data analysis strategy defined

	high-level trigger and offline statistical analysis steps ● Develop unsupervised graph-based ML model optimized to meet inference latency and resource constraints of the high-level trigger system ● Develop method for offline characterization of anomalies to deliver physics results (e.g., through clustering algorithms or semi-supervised strategies)	● Graph-based ML model developed and ready for integration in high-level trigger system
36 m	● Recast the Run 3 L1T anomaly detection algorithm to use Phase 2 particle-based inputs and to additionally perform data compression and reduction tasks for deployment in L1T scouting system ● Develop graph-based transformer models for the three tasks optimized to meet L1T inference latency and resource constraints (e.g., through quantization, pruning, and knowledge distillation)	● Transformer models defined
48 m	● Characterize offline performance (information loss due to compression, event characterization and clustering for anomaly analysis) ● Start co-design transformer models with firmware for deployment in both scouting and production system	● Report on offline performance of the algorithms completed ● First firmware for transformer modules available for testing
60 m	● Complete transformer models with firmware for deployment in both scouting and production system	● Firmware for transformer models deployed

Work Package 4: Education Programmes and Outreach

The education and outreach work package aims to continue development of world-class high-energy scientists and engineers in close collaboration with academic and industry partners to ensure future growth in this field. The activities are organized over two different areas with complementary goals: allowing exchanges by allowing scientists and researchers to come to CERN and work with the project experts; Develop a yearly advanced software Training Programme designed to equip postgraduate students, Ph.D. scholars, and researchers with cutting-edge computing and data science skills.

Task 4.1: Exchange Programmes and Outreach

One of the fundamental ingredients of CERN's recipe for success is the continuous active exchange of experience and knowledge across a broad worldwide community of researchers in many different fields. The NextGen Triggers project requires the implementation of exchange opportunities in the form of dedicated events, project workshops, visiting scientists grants, and specialistic seminars. Visiting scientists' support will allow to host and coordinate the wide community of developers engaged in several of the tasks. The project must also be able to package information about its activities and achievements. For this purpose yearly workshops will be organized for the project, as well as dedicated events for a broad set of audiences, both experts and the general public. The activities in this task will be coordinated by a dedicated Communication and Outreach manager in collaboration with the experts of the technical Work Packages.

Time	Description	Deliverable/Milestone
6 m	Definition and design of the exchange program based on requirements from technical WPs	Outreach and exchange plan
12 m	Execution of the programme	Events and visits according to plan, including a yearly project “all-hands” event
24 m	Yearly updates of planning and execution	Yearly events and visits (including yearly “all-hands” event) Outreach publications on different dissemination channels
36 m		
48 m		
60 m		

Task 4.2: CERN STEAM Programme (CERN Software Training, Education, and Advanced Modules Programme)

The CERN-STEAM Programme is an initiative designed to equip postgraduate students, Ph.D. scholars, and researchers with cutting-edge computing and data science skills, ensuring a vibrant future for the field of research. This comprehensive and immersive educational program focuses on critical areas such as algorithms design, AI, trigger systems, heterogeneous computing, and quantum computing as applied to HEP. Renowned professors and experts from academia and industry will give lectures, seminars, hands-on training, and hackathonsto bridge the skills-gap between academic proficiency and autonomy in developing cutting-edge technologies within the NGT project. The Programme aims to provide an enriching learning experience complementing and building upon the courses taught in established schools and events in the field, through the practical application to CERN experiments' realistic use cases. We will investigate how to make the Programme courses eligible for European Credit Transfer and Accumulation (ECTS) credits.

Collaborations with industry partners will be created to facilitate students' mobility, and training through internships. This will guarantee continuous knowledge transfer between the CERN NextGen Trigger project and industries in the CERN member states.

Time	Description	Deliverable/Milestone
12 m	Develop a year-long program combining novel and existing lectures, seminars, dedicated schools, and "hands-on" training courses in software tools and techniques. Organize pilot trigger software and Data Science training events for the NGT students and researchers.	Program committee established. Skills-gap analysis completed. The initial round of training courses and schools is delivered to NGT students and Fellows.
24 m	Develop a model for joint collaborations with universities for courses with ECTS recognition, an internship program, and technical training courses with industry partners.	The incubation period is over. The second round of training courses and schools is delivered to NGT students and Fellows. The Programme is established as part of the wider CERN education activities.
36 m	The programme runs for three years, providing a comprehensive spectrum of lectures, seminars, and training courses to NGT students and HEP students at large. Upon the success of the pilot events, yearly schools for software development, AI, and Data Science will be organized. Develop a proposal to organize a recurring yearly STEAM Programme.	Grants and funding of activities for doctoral students and researchers. Training, lectures, and seminars delivered to the NGT students and researchers and other interested researchers at CERN. Proposal on how to make the Programme sustainable beyond the project duration submitted.
48 m		
60 m		