Next Generation Triggers 1st technical workshop Report

Roope Niemi
16.12.2024

About the workshop

During 3 days at Prévessin
Over 100 participants in person and online

Monday:

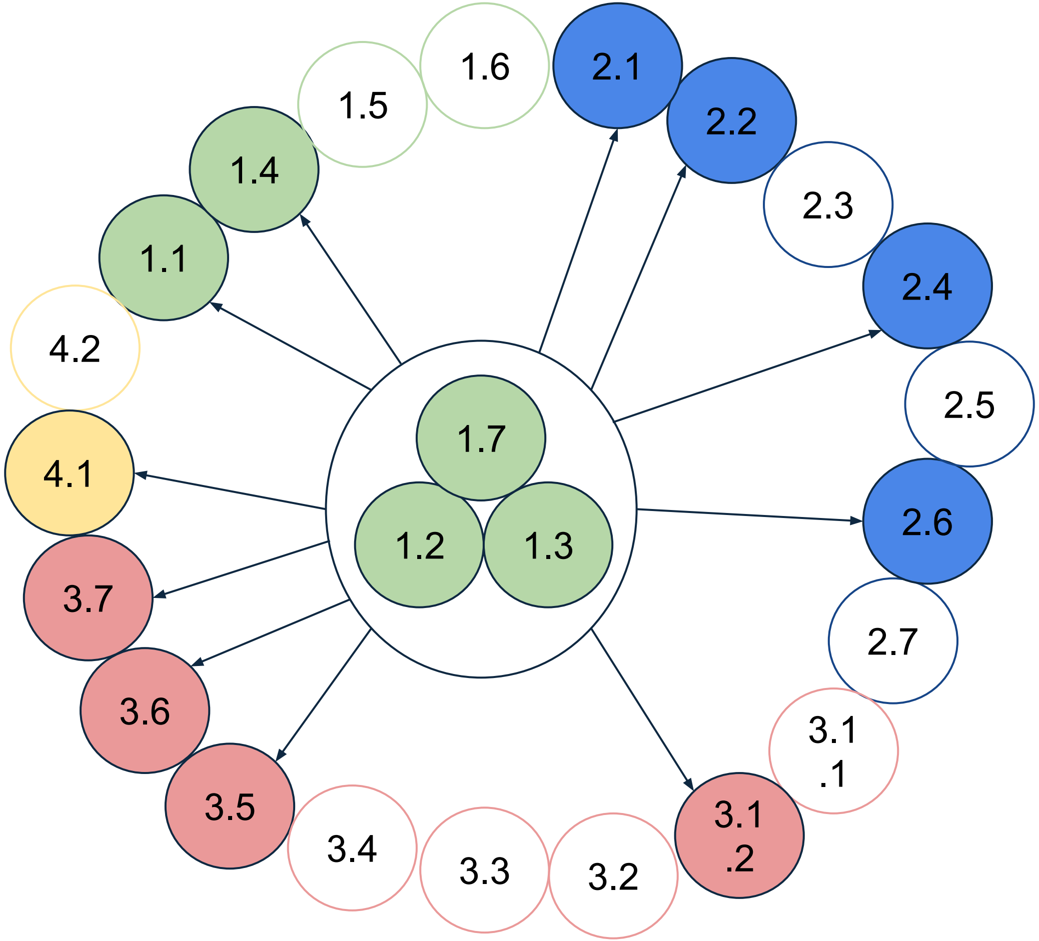
- WP1: Infrastructure, Algorithms and Theory
- WP4: Education Programmes and Outreach

Tuesday & Wednesday:

- WP2: Enhancing the Atlas Trigger and Data Acquisition
- WP3: Rethinking the CMS Real-Time Data Processing

https://indico.cern.ch/event/1421629/overview

We will focus on NGT tasks we are familiar with and/or have discussed collaboration with



1.1 Hardware and services for large scale NN optimisation and training, and physics simulation

Design, procure, deploy and operate the computing infrastructure (hardware and software) and platforms required to support the common tasks in WP1, and tasks in WP2, WP3

Progress:

- Hardware specification and on-premises procurement to be completed by end of 2024
- Initial set of seeding resources available
- Initial set of platforms available (GPUs in GitLab CI, GitHub Actions)

Future:

- Installation and configuration of all on-premises resources and onboard cloud resources
- Complete use case collection and setup benchmark automation
- Deliver a common platform for shared access to project resources
- Deliver an MLOps platform covering the full ML lifecycle

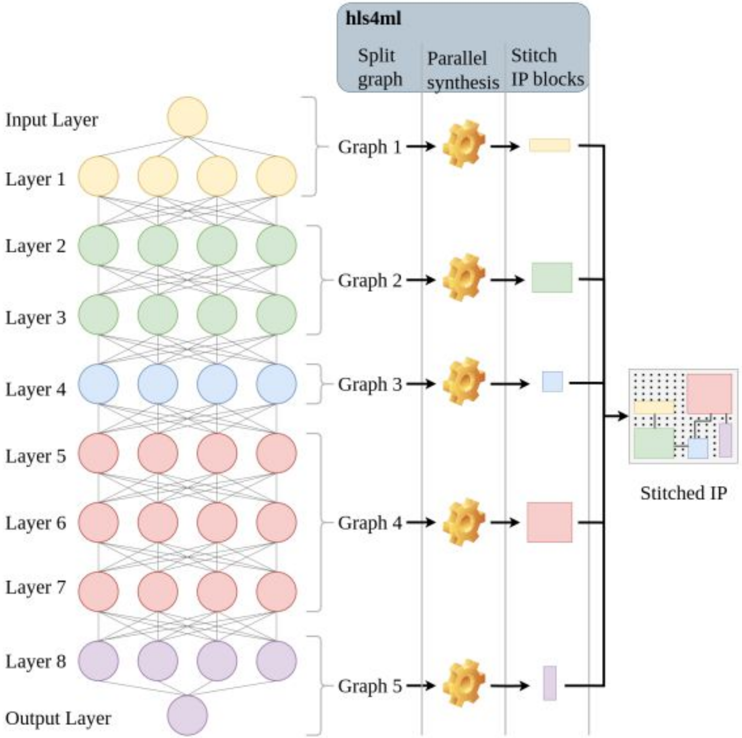
1.2: Fast inference of complex network architectures on LHC online systems

Develop hardware-efficient neural networks and tools for FPGA deployment, while ensuring this work benefits other WPs, as well as FastML/hls4ml community

Given user selected splitting points, partition model graph into smaller, independent subgraphs.

Reduces synthesis and build time, enables efficient handling of larger models, and simplifies debugging

Future: support for larger and more complex models, and for various ML architectures such as transformers, GNNs, seamless workflow for user, simulation of the stitched IP, explore Tree Tensor Networks in collaboration with T1.4



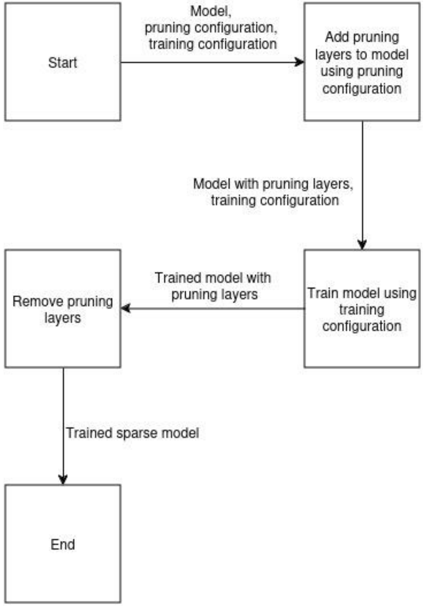
1.3: Hardware-aware AI optimization

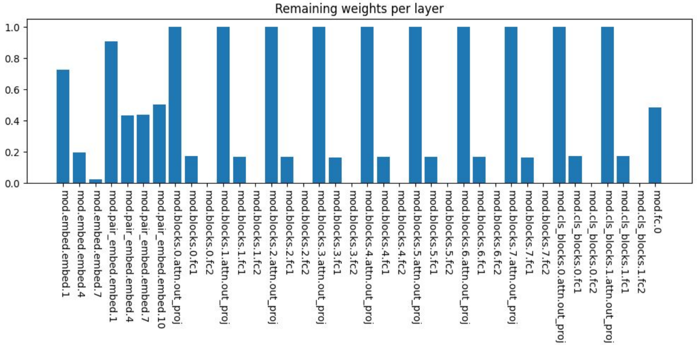
Collect and implement various compression methods. Develop a common, easy to use interface to train and compress models

Current focus on pruning, with YAML based configuration for pruning and training hyperparameters

Initial tests with ResNets and ParT

Future: test more different models, polish library for release and prepare documentation. Also integrate quantization methods, other compression methods

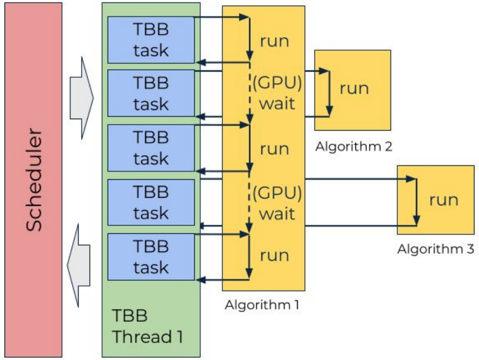

Start
Model,
pruning configuration,
training configuration
Add pruning layers to model using pruning configuration
Model with pruning layers,
training configuration
Train model using training configuration
Trained model with pruning layers
Remove pruning layers
Trained sparse model
End



1.7: Common Software Developments for Heterogeneous Architectures

Efficient use of GPUs and FPGAs in software for HL-LHC.
Harmonized between experiments, with optimization efforts shared
Various topics for R&D:

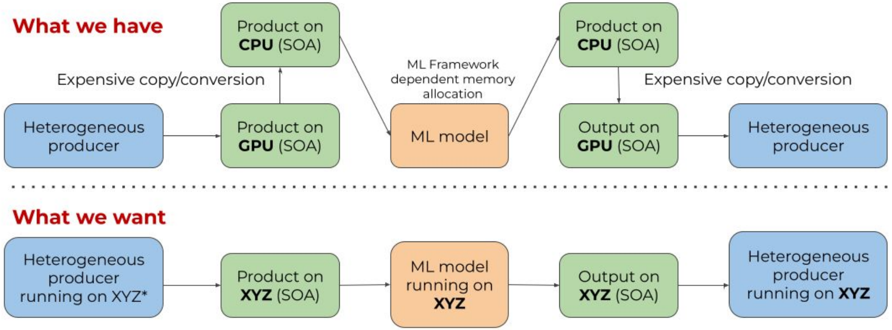
- efficient heterogeneous scheduling
- efficient portable data structures
- efficient accelerator interfaces to ML inference
- common accelerated libraries
- alternative programming languages



Progress:

- prototype of multi-threaded task scheduler
- 2 implementations of data structures with adaptive memory layouts
- CMS local pixel reconstruction fully ported to Julia, performance comparable to C++

Future: More prototyping and benchmarking, start common accelerated libraries and ML inference subprojects



Portable ML model and the *same* memory management as for other heterogeneous algorithms

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

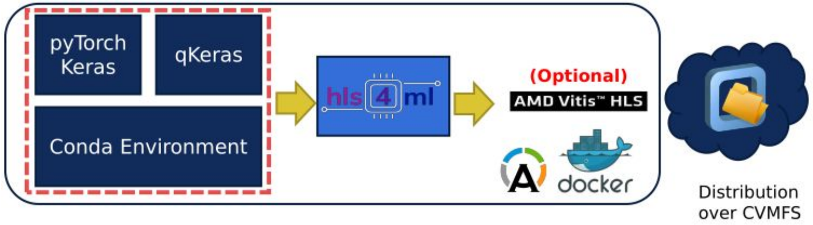
Focuses on implementations within the context of the ATLAS Global Trigger (L0-Global), such as novel ML reconstruction. Data received from the calorimeter and muon systems.

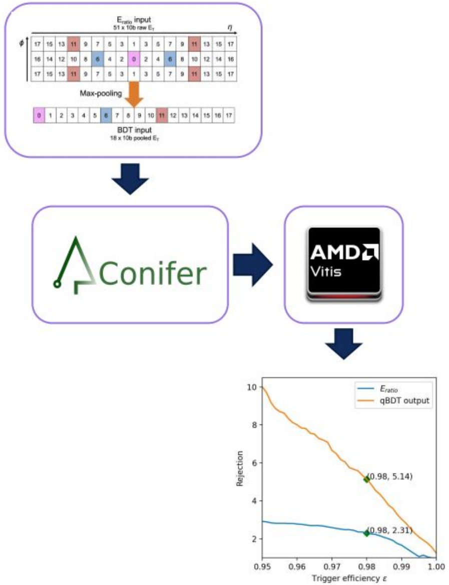
To integrate ML within Global Event Processor, a unified framework required.

Resource extraction: extract performance metrics, such as resource plotting, physics performance, plotting scripts and CLI for cross-platform comparisons

Evaluated new technologies

- BDT implemented in HLS, showing promising latency
- Also CNN implementation with hls4ml in progress





17	15	13	11	9	7	5	3	1	3	5	7	9	11	13	15	17
16	14	12	10	8	6	4	2	0	2	4	6	8	10	12	14	16
17	15	13	11	9	7	5	3	1	3	5	7	9	11	13	15	17

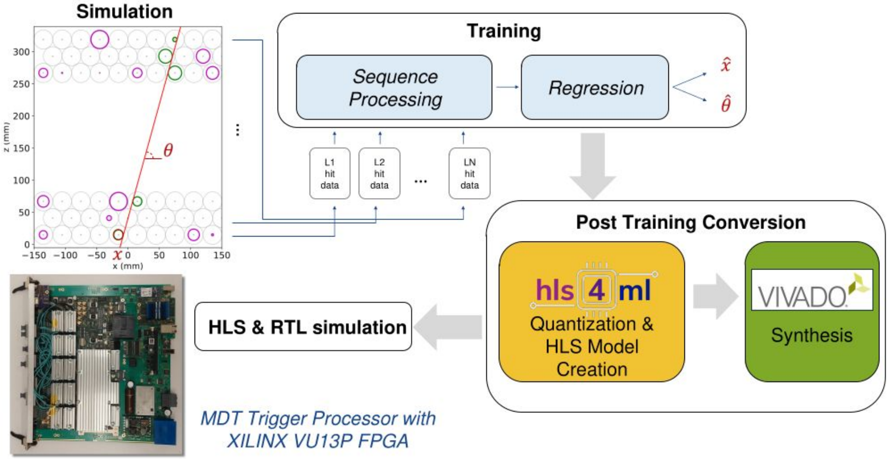
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17

2.2: Enhancing the L0 Muon Trigger ATLAS Work Package

Focus on improving robustness of L0 Muon trigger, and extending to new signatures such as long-lived particle decays

Investigate uses of ML for L0 Muon trigger, promising results for pattern recognition and momentum estimation

Priorities for 2025, develop algorithms further. Shift towards implementations for HW. Use computation resources from WP1.



MDT Trigger Processor with XILINX VU13P FPGA

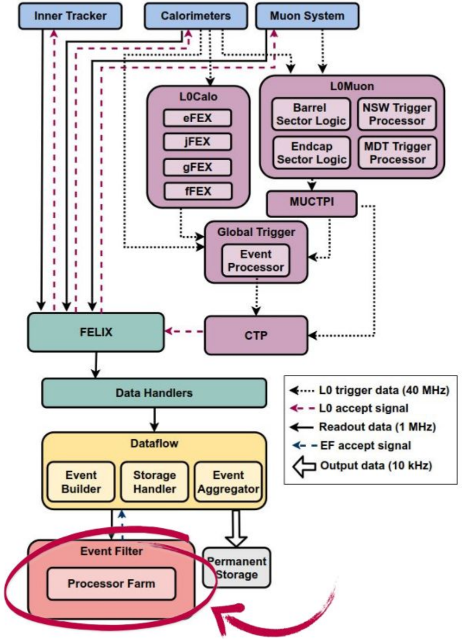
2.4: Event Filter Tracking

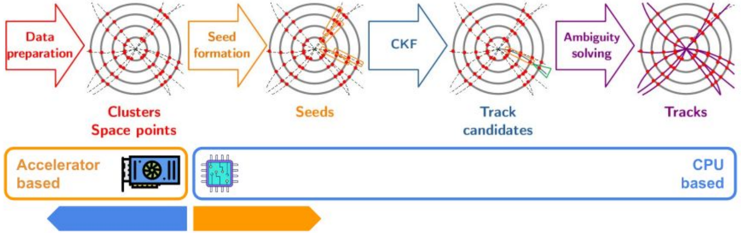
Goal: Development of an algorithmic solution for the ATLAS EF track reconstruction

- Exploring optimal classical numerical and ML techniques, e.g. GNNs
- Deployment on the most suitable hardware architecture: FPGA, CPU, GPU

Progress:

- FPGA implementation of each track reconstruction step. Need to find optimal boundary between FPGA and CPU





Sliding window which can be optimized for FPGA solution

2.6 Common Event Filter Infrastructure

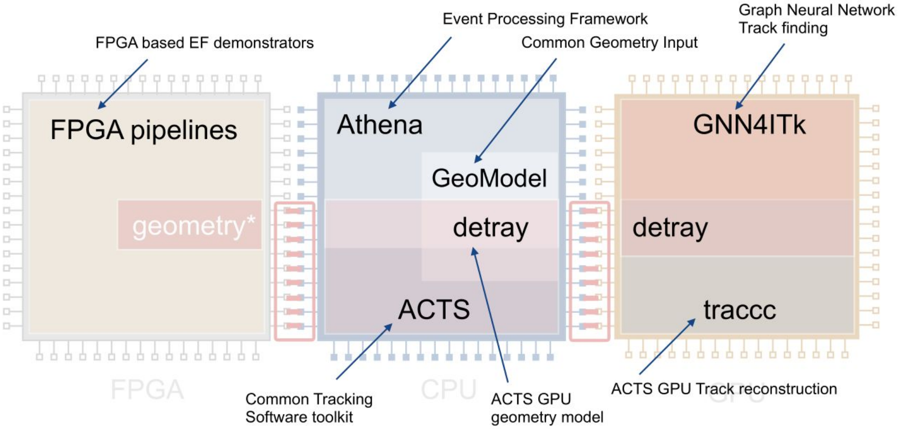
Goal:

- Provide infrastructure for the various ATLAS EF prototypes within ACTS
- Improve integration of GPU/CPU/FPGA demonstrators
- Help with interfacing ML/AI pipelines with ACTS
- Optimize data structures for heterogeneous pipelines
- Optimize throughput of candidate CPU/GPU EF applications

Progress:

- Work on improving ACTS track reconstruction chain, in cooperation with 2.4
- Work on optimizing of the GPU pipeline
- Work on next generation geometry support

Collaborating with 1.7, 2.4, 2.5



FPGA based EF demonstrators
Event Processing Framework
Common Geometry Input
Graph Neural Network Track finding
FPGA pipelines
geometry*
Athena
GeoModel
detray
ACTS
GNN4ITk
detray
traccc
FPGA
CPU
GPU
Common Tracking Software toolkit
ACTS GPU geometry model
ACTS GPU Track reconstruction

3.1.2 Efficient data structures for heterogeneous event reconstruction

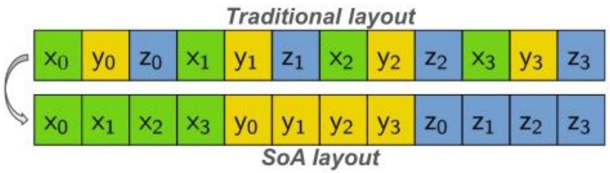
Goal:

- Efficient memory access patterns
- Seamless integration with ML models, remote offload through message passing
- Flexibility, maintainability

Standardizing data structures is crucial. Must be compatible with C++20

Flexible SoA Composition: Dynamic data requirements, carry only necessary data through the processing pipeline. Minimization of data copying, efficiently use memory and bandwidth.

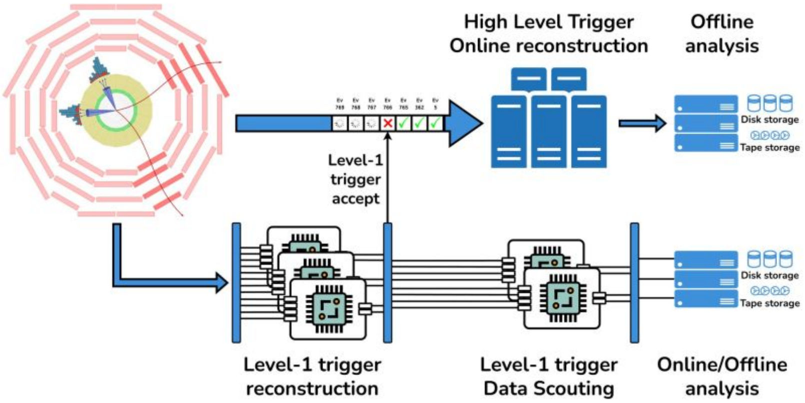
Collaborating with 1.7, EP-SFT and CMS ML group


Traditional layout
x_0 y_0 z_0 x_1 y_1 z_1 x_2 y_2 z_2 x_3 y_3 z_3
x_0 x_1 x_2 x_3 y_0 y_1 y_2 y_3 z_0 z_1 z_2 z_3
SoA layout

3.5 L1 scouting for HL-LHC

Data scouting studies signatures evading the standard trigger chain, with dedicated FPGA-based boards that collect L1T objects

They created a demonstrator for scouting data acquisition, and are looking into extending the demonstrator with AMD AI engines, FPGAs and GPUs

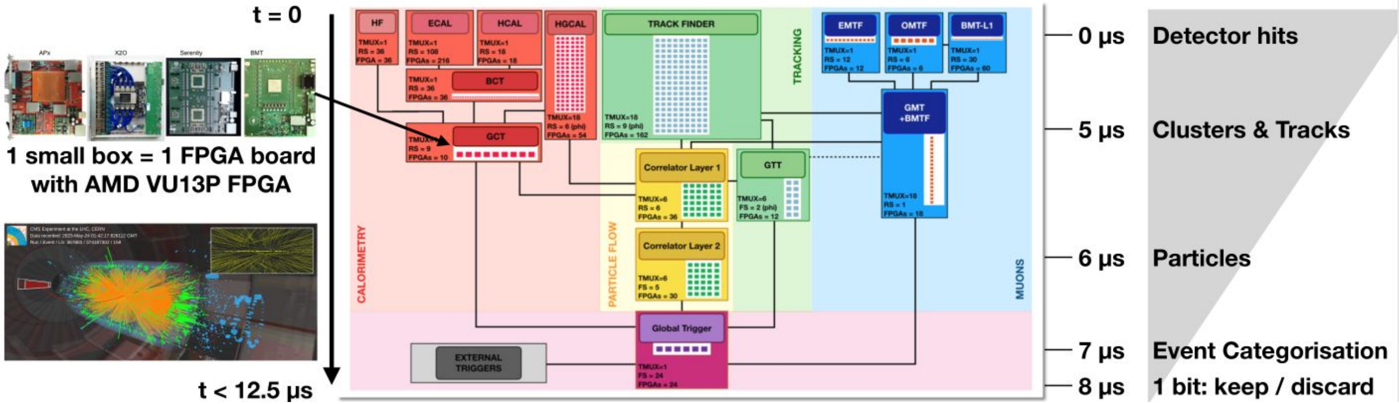


3.6: Practical real-time AI for Level 1 Trigger and L1 Scouting

Goal: 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

Progress:

- Improved and consolidated jet tagging for correlator trigger, with addition of multi-class and pT regressing model
- Improved HGCal cluster ID for better Particle Flow and electron reconstruction
- Adopted MLOps, using MLflow and GitLab CI



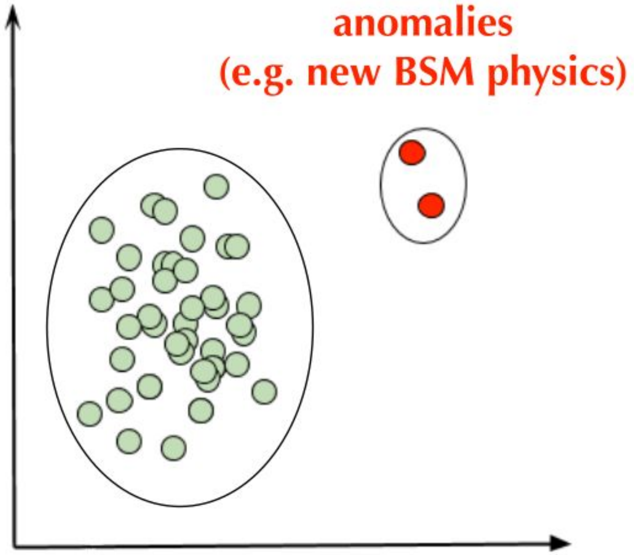
1 small box = 1 FPGA board with AMD VU13P FPGA

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

Goals:

- generalize new physics searches to a large variety of BSM models at once
- Identify anomalies in data
- Learn normal behaviour with neural networks such as autoencoders

Anomaly detection trigger based on autoencoders has already been integrated in the system via hls4ml.
The team is looking into improving their training pipeline using contrastive learning and compression techniques, and explore new neural network architectures.



Outreach (WP4.1) and Education Programmes (WP4.2)

4.1: Ensure continued skill development of scientists and engineers able to combine domain-specific knowledge of HEP with data science and AI. Promote exchanges by allowing scientists and researchers to come to CERN and work with project experts. Allow project members to visit external institutes and companies.

4.2: Equip postgraduate students, Ph.D. scholars, and researchers with cutting-edge software skills. Focus on algorithms, AI, trigger systems, advanced computing as applied to HEP.

We have talked about organizing a workshop/tutorial on tools developed in 1.2 and 1.3

Other tasks

WP1

1.4 **goal**: Develop and apply quantum-inspired methodology, in particular Tensor Network algorithms, to simulate quantum many-body problems

Progress: First results on Tensor Networks and Quantum ML analysis. Ongoing projects on real-time dynamics of High-Energy Physics and classical simulation within GPUs architectures

1.5 **goal**: Port and optimize current event-generation codes and higher-order perturbative calculations to state-of-the-art and future hardware architectures, particularly GPUs

Progress: First release for acceleration of leading-order processes on GPUs and vector CPUs. Work on integration with CMS.

Provide benchmarking support with lattice QFT codes guiding hardware procurement and commissioning for HPC hardware.

1.6 **goal**: Identify BSM scenarios and signatures to be used as benchmarks for the assessment of new-generation triggers performance, in collaboration with the experiments. Formal start in 2025

WP2

2.3 **goal**: Optimize the readout performance and address bottlenecks in the system

Progress: First specification of CPU requirements. Identified candidate CPUs, identified/compared network interface cards. Compared different types of servers. New cost-saving architecture devised and costed. Progress in the development of the netio3 library.

2.5 **goal**: Fully exploit the extended coverage of the L0 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. Develop novel ML based reconstruction techniques to improve on existing classical algorithm chain

Progress: Understand where to concentrate effort on traditional approach. Working on migrating existing software to ACTS, new geometry representations, code optimization, and novel ML reco techniques + associated tools

2.7 **goal**: Exploit the enhanced functionality and performance of the novel tracking approaches, including those developed in 2.4 and 2.5, to extend the physics potential of ATLAS and develop novel algorithmic approaches to efficiently search for non-standard particle signatures

Progress: First processes and signatures identified, simulation of samples initialized. Investigating physics potential of TLA and techniques such as anomaly detection.

WP3

3.1.1 **goal**: Modernize the traditional CMS Phase-2 reconstruction 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.

3.2 **goal**: Extend the CMS data processing framework to become capable of adapting to different network topologies to leverage remote accelerators, with little or no modification to the core code.

Progress: Implementation of a client-server, multithreaded, distributed test application, based on CMSSW, leveraging high-speed host-to-host or shared memory communication.

3.3 **goal**: increase the maximum trigger available by reducing the size of events saved by CMS

Progress: Produced a report that illustrates the impact of RAW data compression and of their replacement with low-level reconstructed quantities

3.4 **goal**: Design accelerated calibration workflows that achieves at the High-Level Trigger (HLT) the same level of accuracy as that of offline reconstruction, by introducing online data buffering and exploiting predictive AI techniques in the calibration.

Progress: Surveyed and identified initial set of candidate conditions for NGT demonstrator. Initial version of the test harness for evaluation of impact of different conditions. By end of 2024 create report illustrating the current calibration workflows, evaluate the impact