
CERN
NexTGen
Next Generation Triggers

Next Generation Triggers

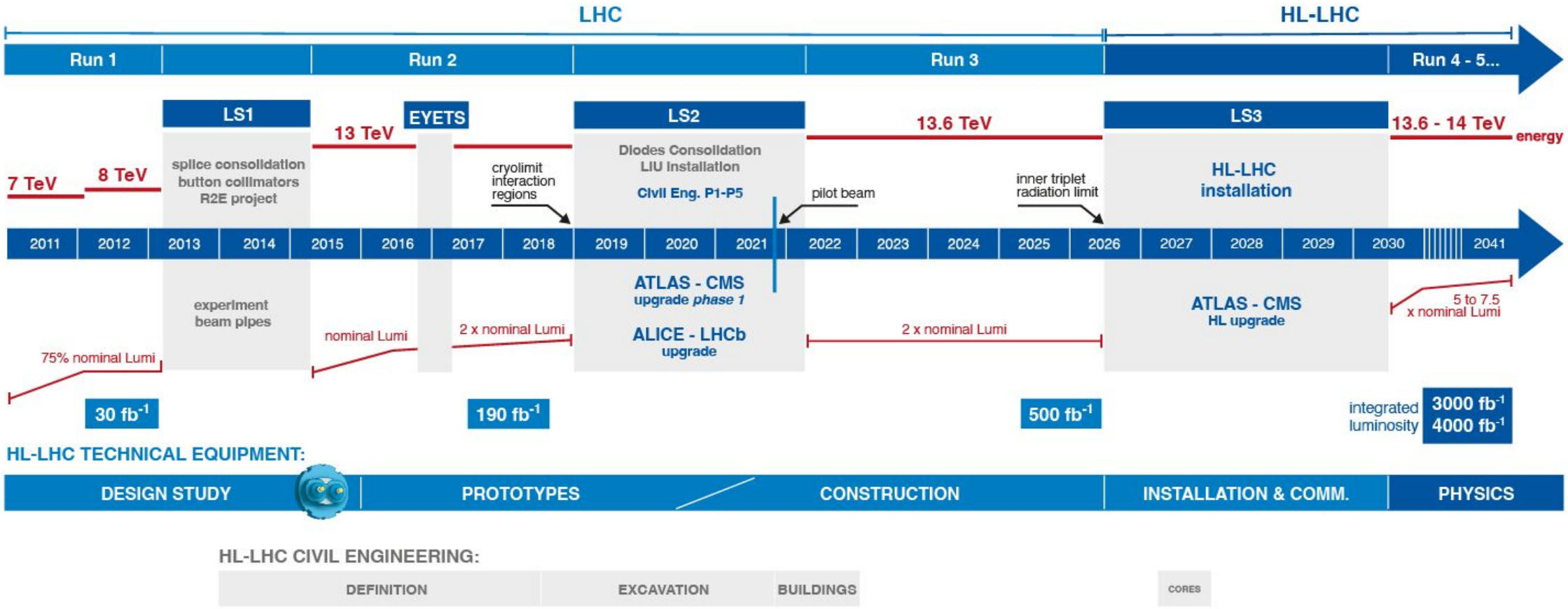
Sebastian Dittmeier
Physikalisches Institut - Heidelberg University
on behalf of the Next Generation Triggers Project

EPIGRAPHY Kickoff school and workshop
12.01 – 15.01.2026
Lisbon, Portugal

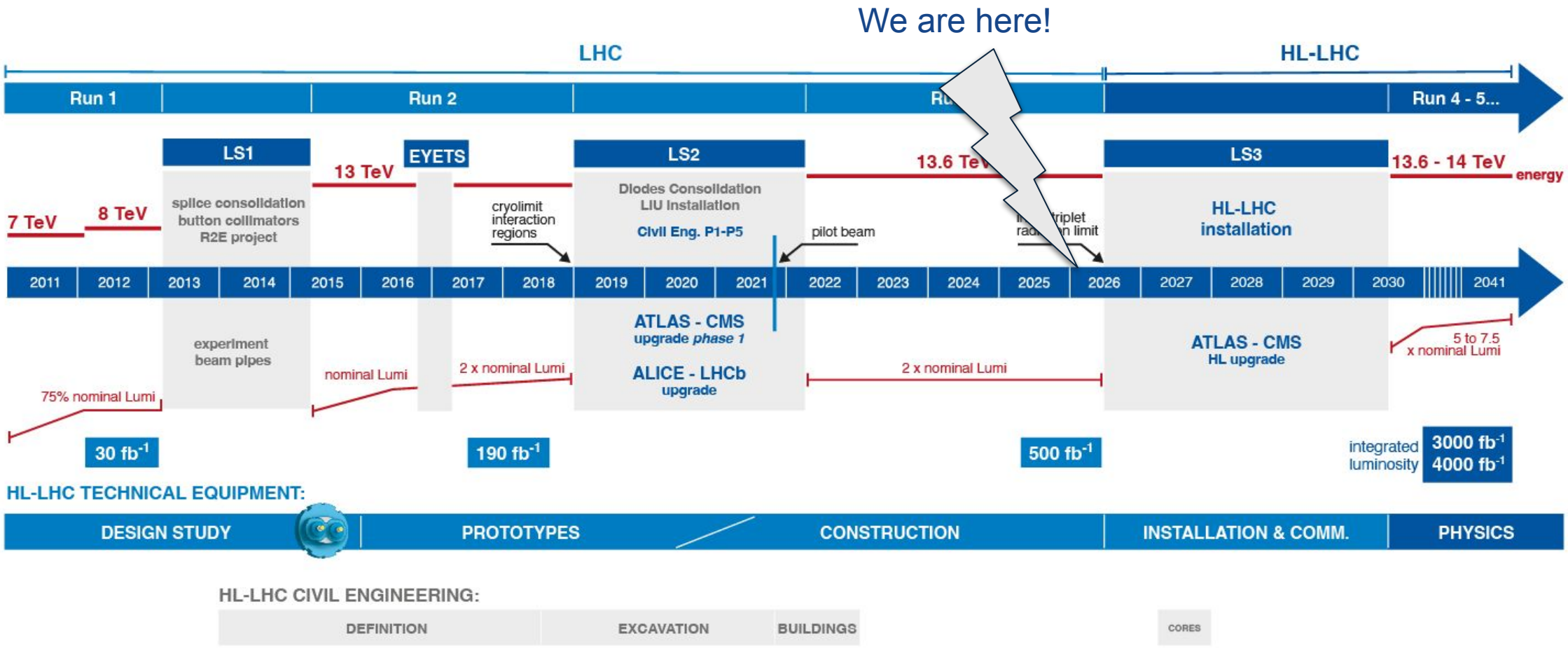
CERN | NexTGen Next Generation Triggers

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

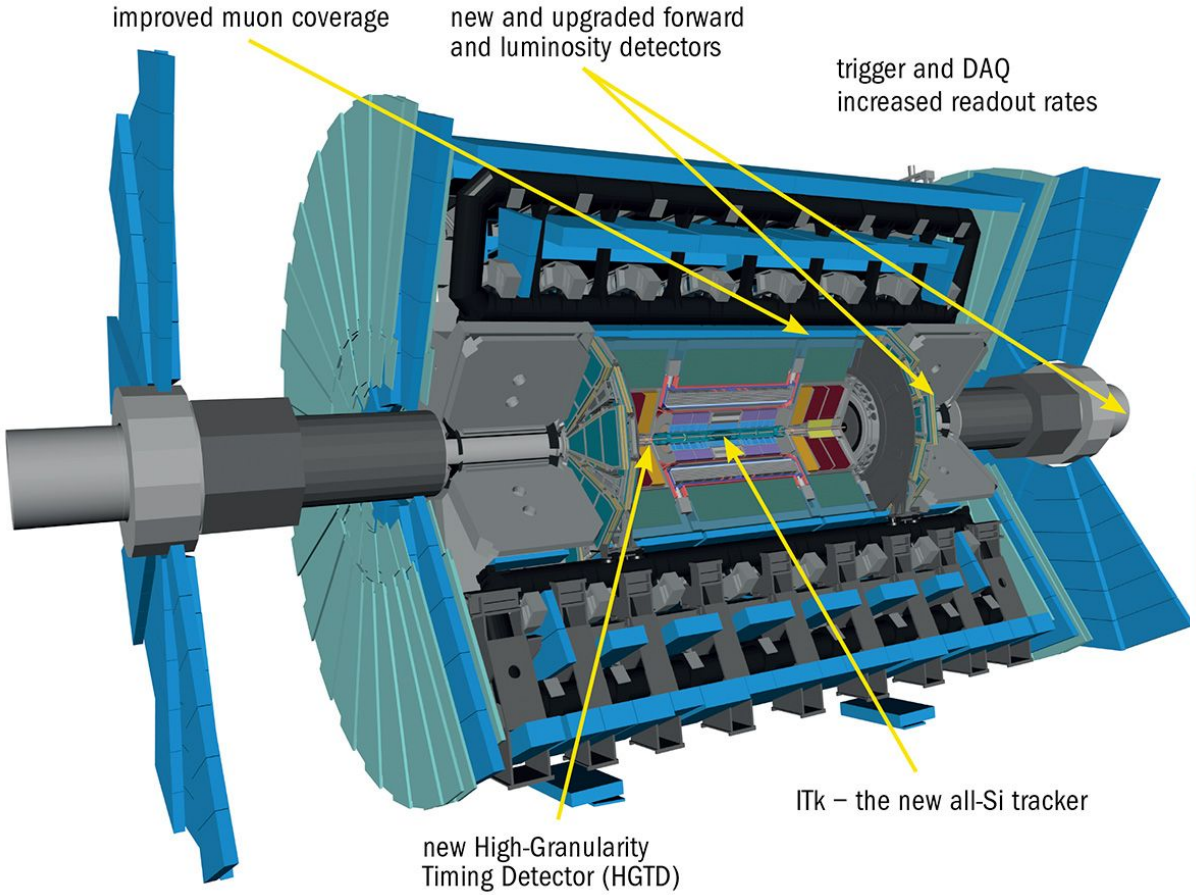
High Luminosity Large Hadron Collider



High Luminosity Large Hadron Collider

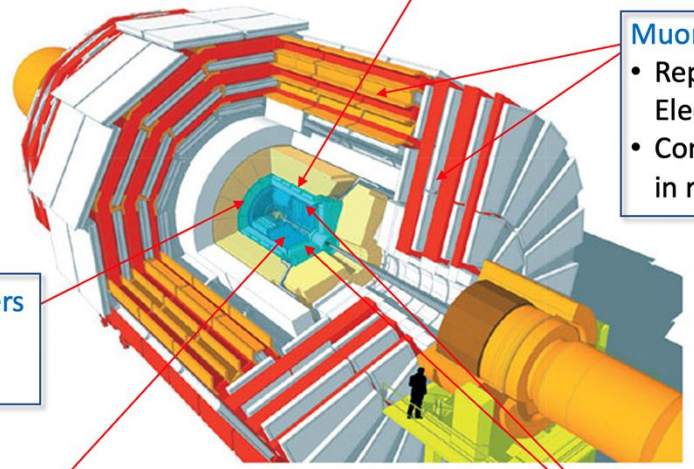


ATLAS and CMS upgrades



Trigger/HLT/DAQ
- Track information in L1-Trigger
- L1-Trigger: 12.5 μs latency – output 750 kHz
- HLT output 7.5 kHz

Barrel ECAL/HCAL
- Replace FE/BE electronics
- Lower ECAL operating temp. (8 °C)



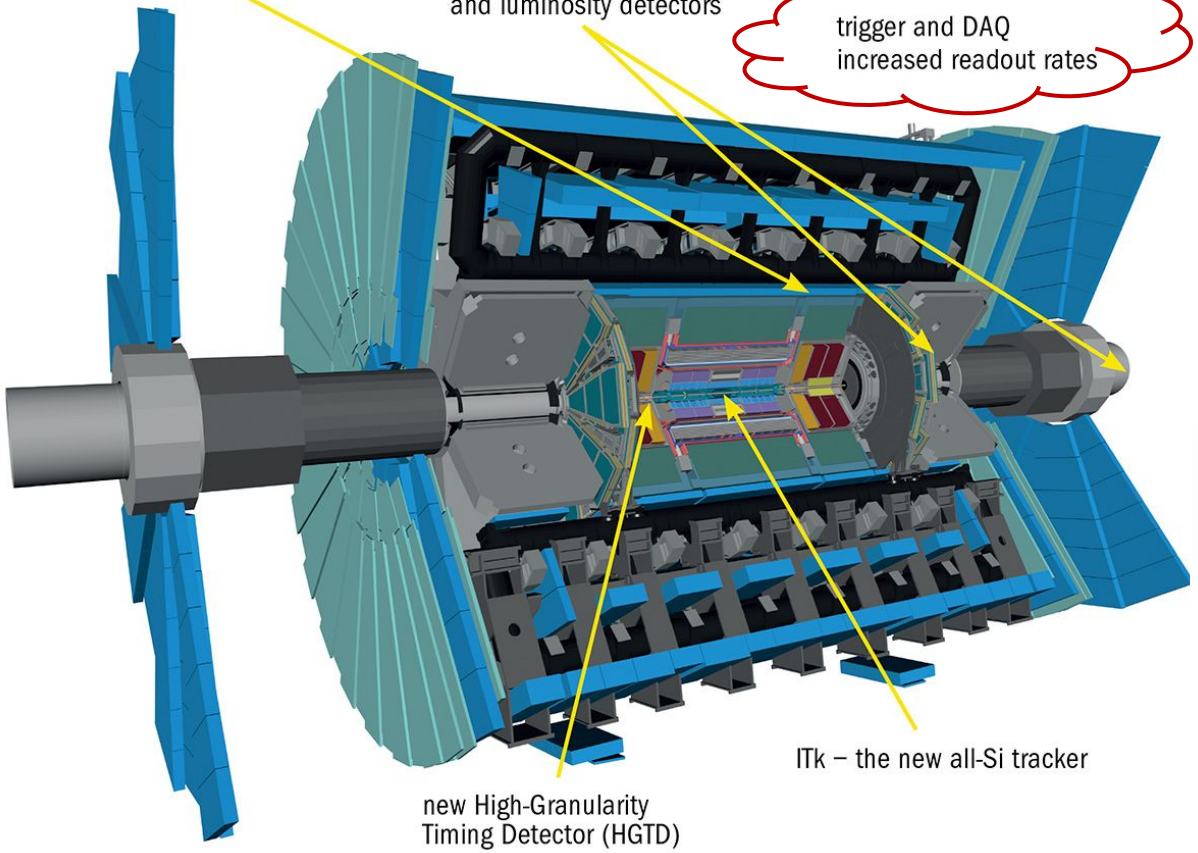
Muon Systems
- Replace DT & CSC FE/BE Electronics
- Complete Muon coverage in region $1.5<\eta<2.4$

New Endcap Calorimeters
- High granularity
- 3D capable

New Tracker
- Rad. tolerant – high granularity – significant less material
- 40 MHz selective readout (p_T>2 GeV) in Outer Tracker for L1 -Trigger
- Extended coverage to η=4

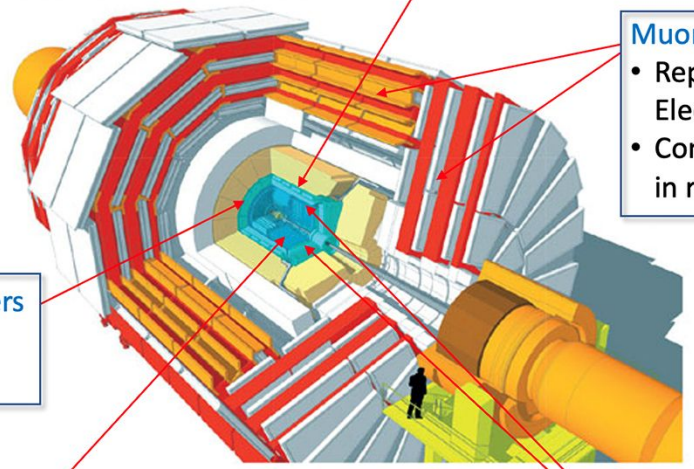
New Precision Timing Detector
- Barrel: Crystal +SiPM
- Endcap: Low Gain Avalanche Diodes

ATLAS and CMS upgrades



Trigger/HLT/DAQ
- Track information in L1-Trigger
- L1-Trigger: 12.5 μs latency – output 750 kHz
- HLT output 7.5 kHz

Barrel ECAL/HCAL
- Replace FE/BE electronics
- Lower ECAL operating temp. (8 °C)



Muon Systems
- Replace DT & CSC FE/BE Electronics
- Complete Muon coverage in region $1.5<\eta<2.4$

New Endcap Calorimeters
- High granularity
- 3D capable

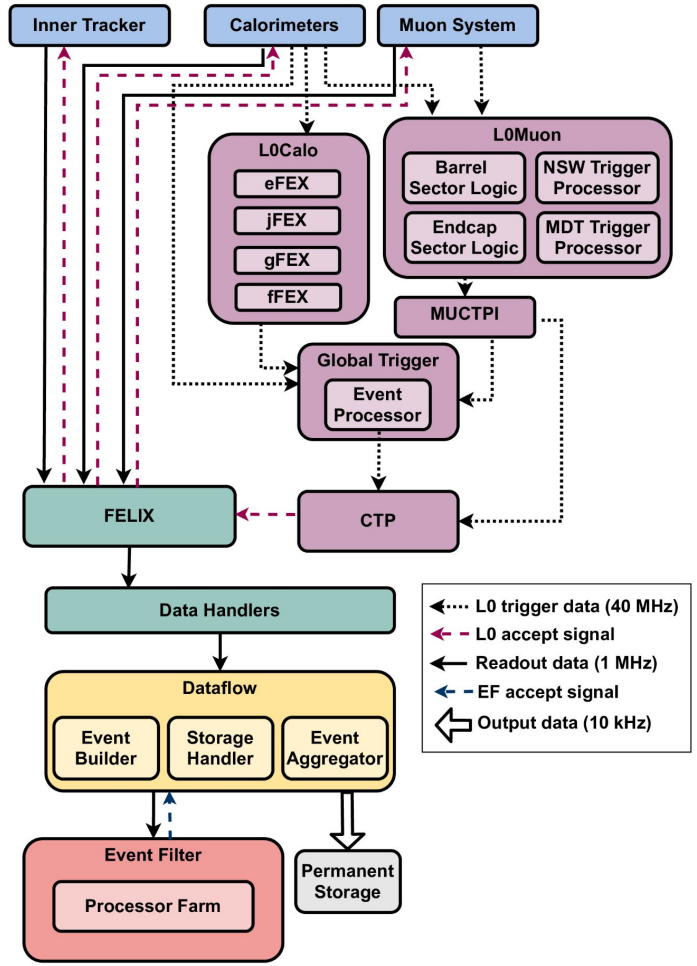
New Tracker
- Rad. tolerant – high granularity – significant less material
- 40 MHz selective readout (p_T>2 GeV) in Outer Tracker for L1 -Trigger
- Extended coverage to η=4

New Precision Timing Detector
- Barrel: Crystal +SiPM
- Endcap: Low Gain Avalanche Diodes

ATLAS and CMS trigger upgrades

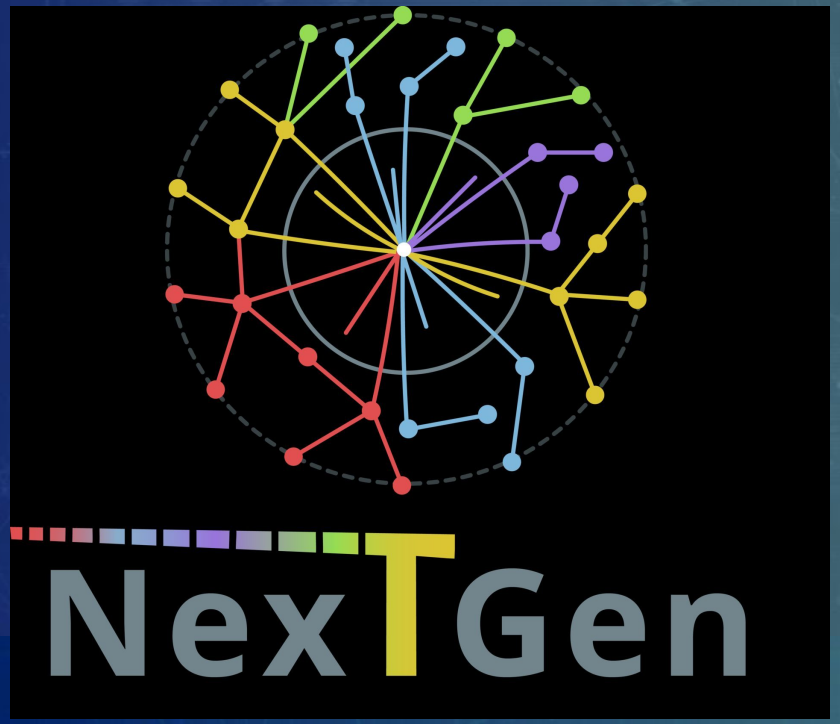
- **Challenge**: Keep the same physics acceptance up to 200 pile-up interactions (currently ~ 60)
- Upgraded sub-detectors backends with newer technologies
  → afford higher L1 Latency (2.5-3.8 μs →10-12 μs)
  and L1 rate up to 750-1000 kHz
  → **Still can't afford to read everything**
- **Hardware Triggers**
  - ATCA boards and AMD FPGAs in many flavors (eg VU13P)
  - ATLAS: new Global Trigger | CMS: new Track Trigger @ L1
- **Software Triggers**
  - ATLAS: CPUs + potential accelerators
  - CMS: GPUs since Run3, more offloading planned

ATLAS Phase II TDAQ system



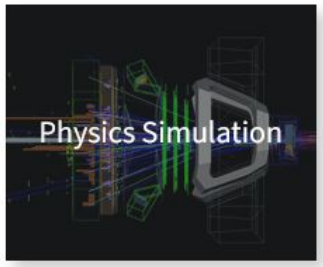
But what if New Physics doesn't look as expected and it is buried under the bulk of the background events that we are still throwing away?

But what if New Physics doesn't look as expected and it is buried under the bulk of the background events that we are still throwing away?



Next Generation Trigger Project

- **Goal:** Develop new strategies for more sophisticated event processing and selection beyond the currently projected HL-LHC upgrade scope, with particular emphasis on **Machine Learning**.
- **Five years project (2024-2028), enabled by an external donation, combining**
  - ATLAS and CMS collaborations
  - CERN's Theory & IT departments
  - CERN's Experimental Physics Software group
- **Opportunities for wider R&D:**
  - Heterogeneous computing
  - FastML tools
  - Workflows for ML in trigger systems
  - Event generators
  - …

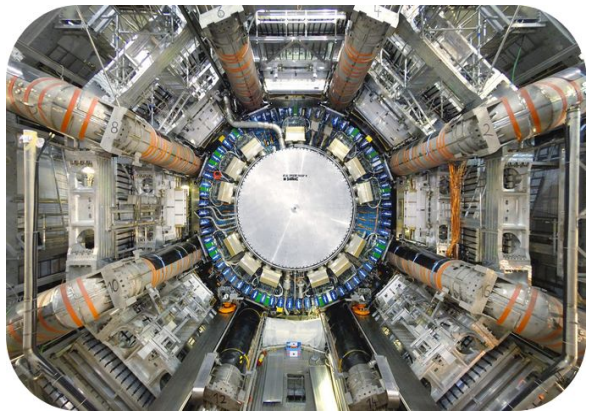


Key Objectives

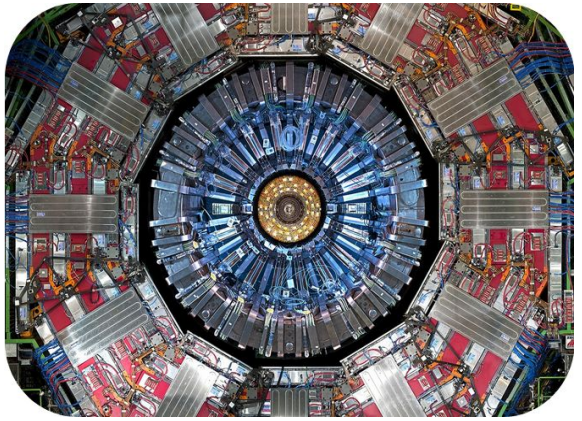
- **Experiment-specific R&D**
  - Requirements defined by the experiments
  - Benefitting from common R&D and training
- **Common R&D on technology aspects**
  - ML and classical algos: invent, optimize, benchmark
  - FPGA deployment, GPU deployment
- **Community aspects**
  - Training experts to stay in the field
  - Educating the next generation of experts
  - Results are open: open access, open source, including training

Next Generation Trigger Activities

WP1: Infrastructure, Algorithms and Theory



WP2: Enhancing the ATLAS Trigger and Data Acquisition



WP3: Rethinking the CMS Real-Time Data Processing

WP4: Education Programmes and Outreach

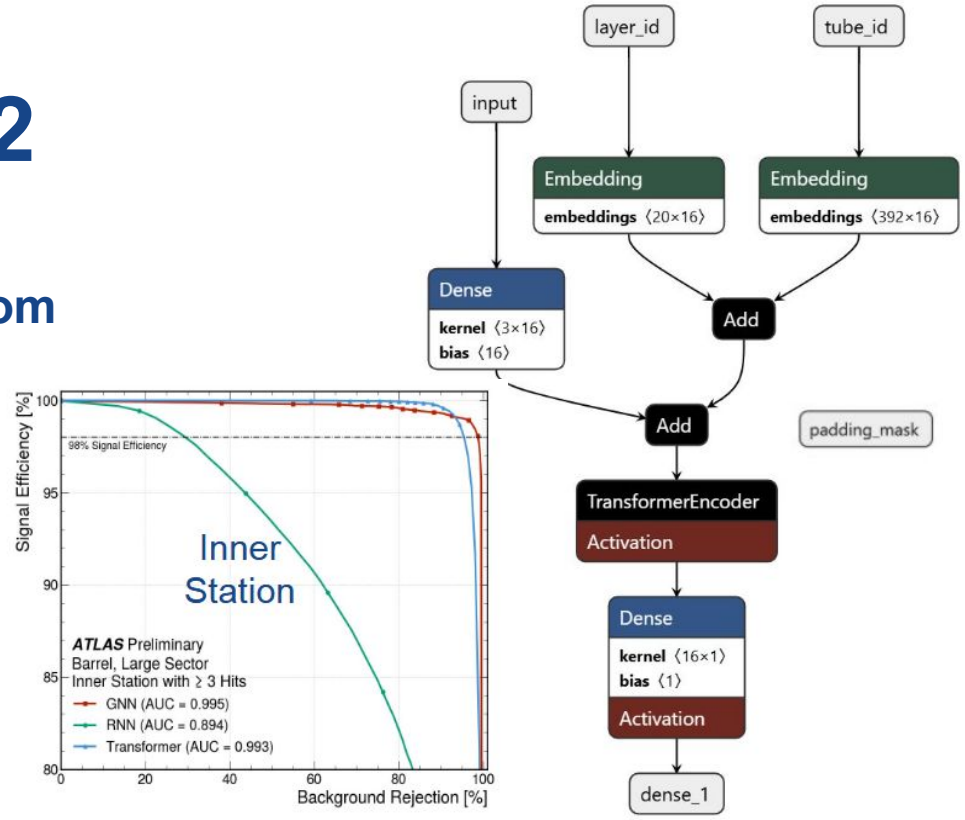
WP2 & WP3

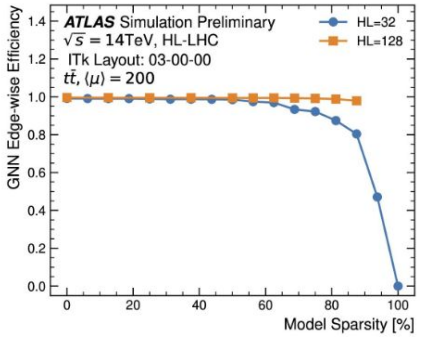
1. Optimal Real-Time Event Selection in the **Global Trigger system**
2. Enhancing the **Level-0 Muon Trigger**
3. High Throughput **Data-Collection**
4. **Event Filter Tracking**
5. Optimized **Event Filter Muon** Trigger Selection
6. Common **Tracking Event Filter** infrastructure
7. Enhanced **Reconstruction** for **Higher Level Event Filtering**

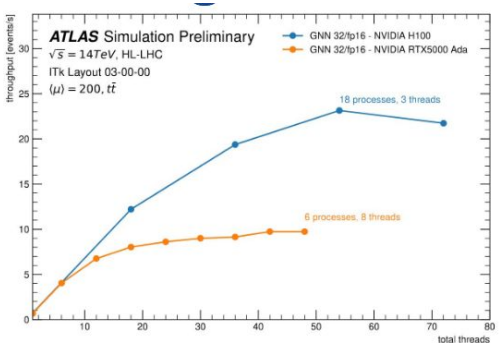
1.1. **The Real-time Reconstruction Revolution (R^3 – Rcube)**
1.2. **R^3** optimized data structures for **heterogeneous platforms**
2. Evolving the **CMS experiment software** into a **client-service distributed application for HLT**
3. **Reduction** of the **RAW data size** for **HLT**
4. **Optimal calibration for HLT**
5. **L1 scouting** for HL-LHC
6. **Practical real-time AI for Level 1 Trigger and L1 Scouting**
7. **L1 scouting data compression** for efficient data acquisition and **anomaly detection**

Selection of Highlights from WP2

- **Transformer-encoder model to discriminate muon hits from background hits for L0muon**
  - Using positional encoding to learn spatial structure of event and the hit locations with respect to each other
  - Aim to decrease model size (8,400 parameters, per station) and reduce input sequence length
  - Comparing with GNN and RNN
- **GNN-based tracking on GPU for EF Tracking**
  - Speed optimization
  - Efficient GPU utilization and negligible overhead
  - Structured pruning investigations to reduce model size

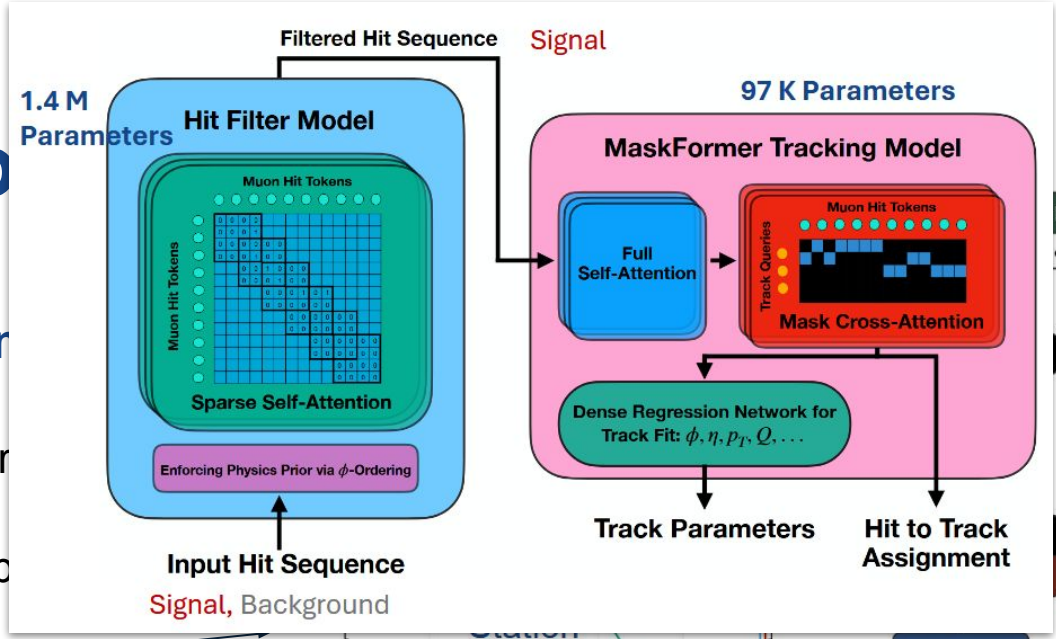


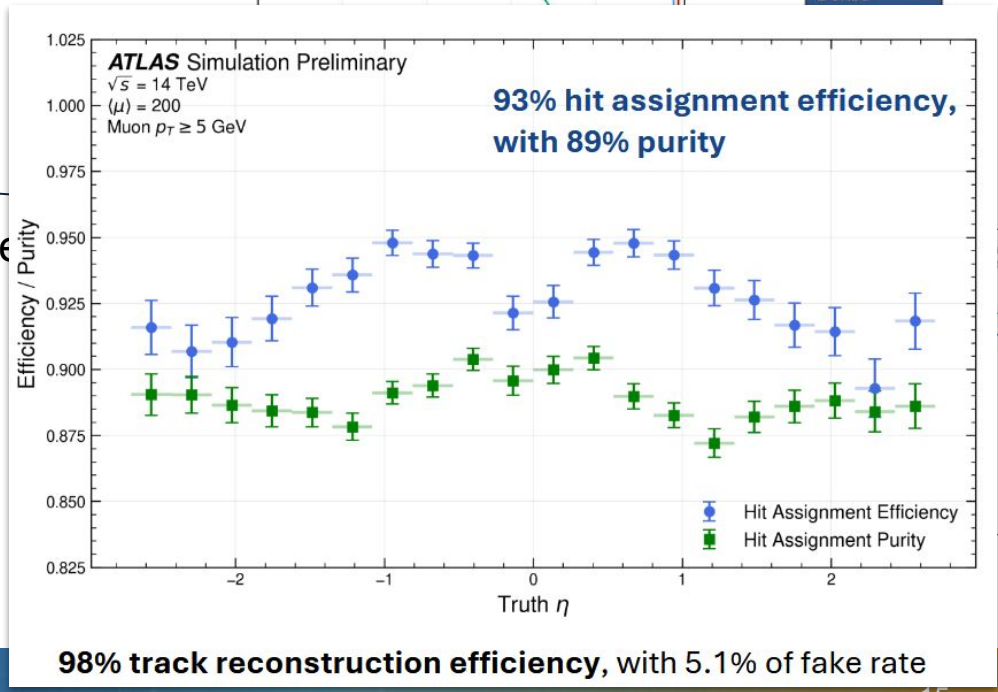




Selection of Highlights fro

- **Transformer-encoder model to discriminate r background hits for L0muon**
  - Using positional encoding to learn spatial structur and the hit locations with respect to each other
  - Aim to decrease model size (8,400 parameters, p and reduce input sequence length
  - Comparing with GNN and RNN
- **GNN-based tracking on GPU for EF Tracking**
  - Speed optimization
  - Efficient GPU utilization and negligible overhead
  - Structured pruning investigations to reduce model size
- **End-to-End ML Muon Tracking**
  - Using MaskFormer open-source model by Meta AI
  - Goal: Assign good hits to tracks and produce an initial track parameter estimate

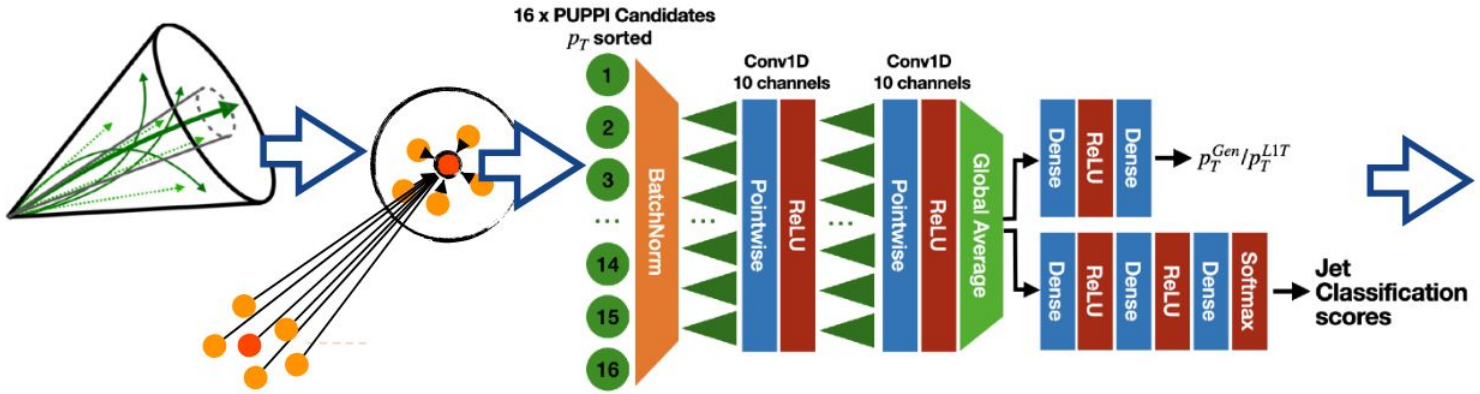


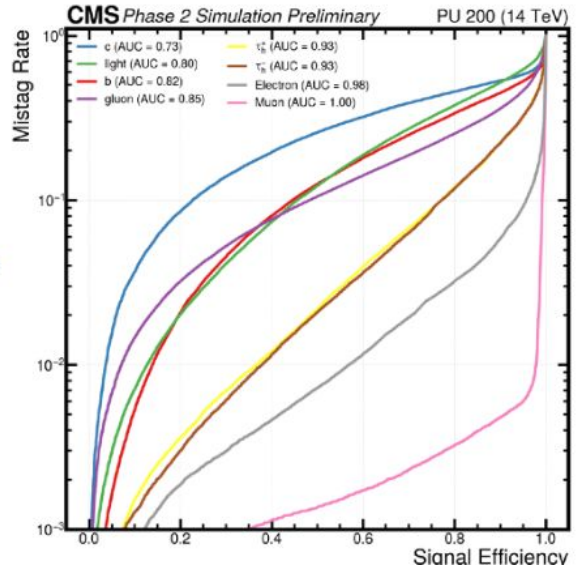


Selection of Highlights from WP3

- **Jet tagging**
  - Small Deep Sets Neural Network trained using Quantization Aware Training and deployed to FPGA with hls4ml
  - Fitting nicely into hardware alongside jet reconstruction firmware

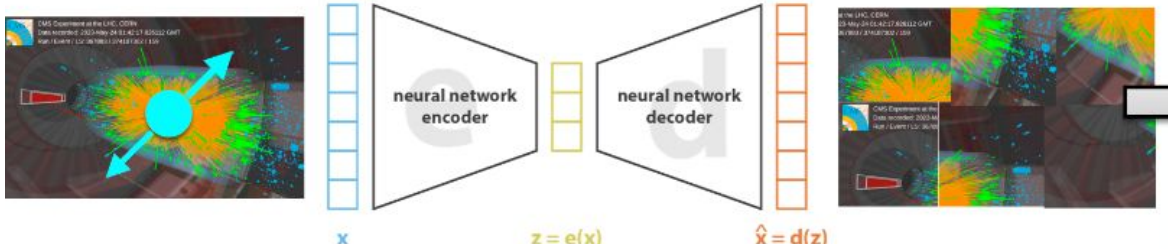
Component	Latency (μs)	LUT	FF	DSP	BRAM
Preprocessing (incl. sort)	0.10	6%	4%	1%	0.6%
Jet Tag NN	0.19	7%	5%	14%	1%
Jet Reco	0.74	9%	7%	5%	0%
Total System	1.01	25%	15%	19%	1%

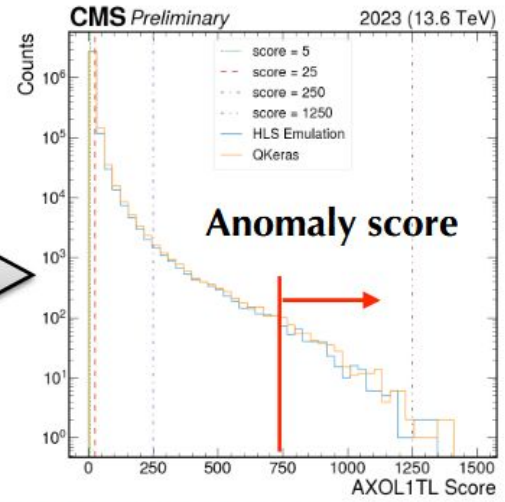




Selection of Highlights from WP3

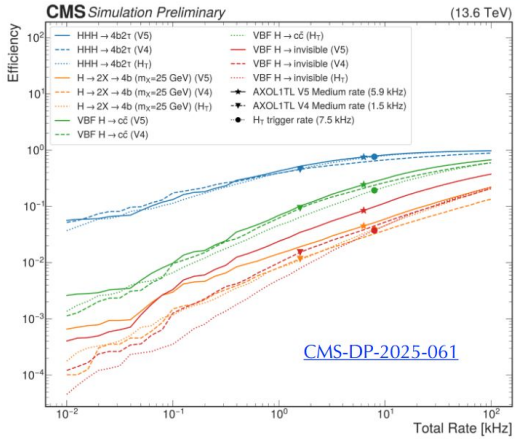
- **Anomaly Detection**
  - Aim of improving AXOL1TL → a fast autoencoder approach for CMS L1T to trigger on rare, unusual events in real time





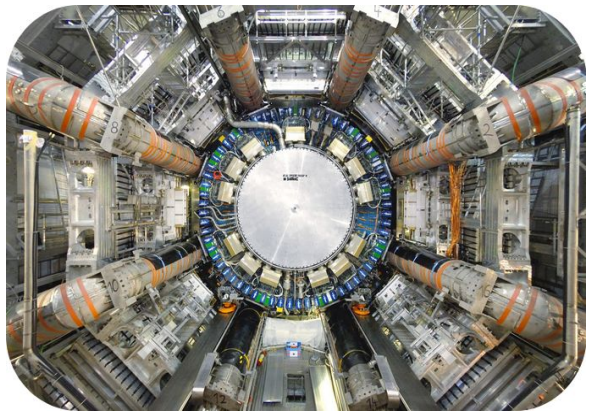
save only anomalous events

  - Model developments using Contrastive Learning: a self-supervised learning (SSL) technique that aims to learn representations by comparing similar and dissimilar samples (called “augmentations”)
  - Same latency and resource usage
  - Improvements observed in efficiency (v5 vs v4)



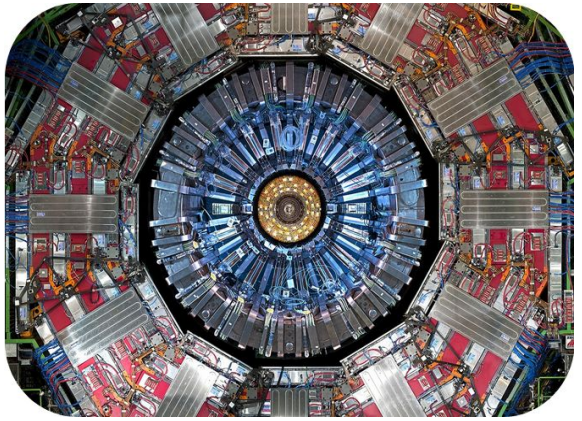
Next Generation Trigger Activities

WP1: Infrastructure, Algorithms and Theory



WP2: Enhancing the ATLAS Trigger and Data Acquisition

see talk by David



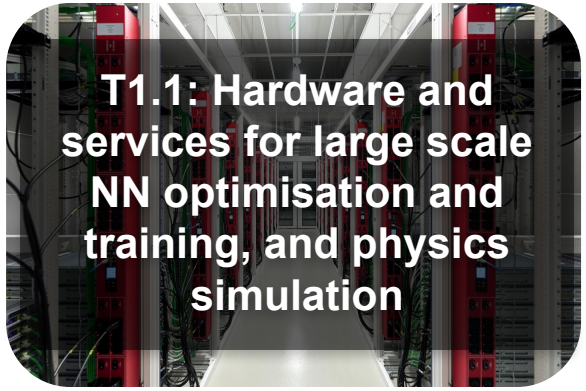
WP3: Rethinking the CMS Real-Time Data Processing

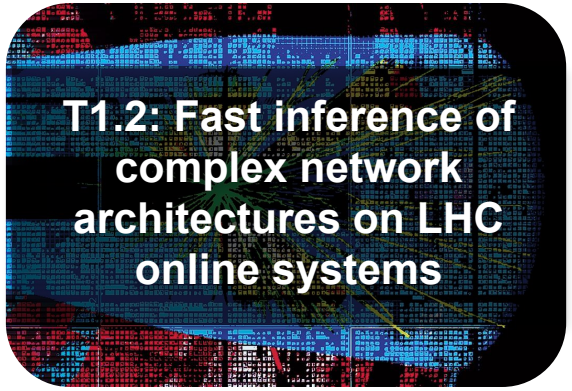
see talk by Emyr

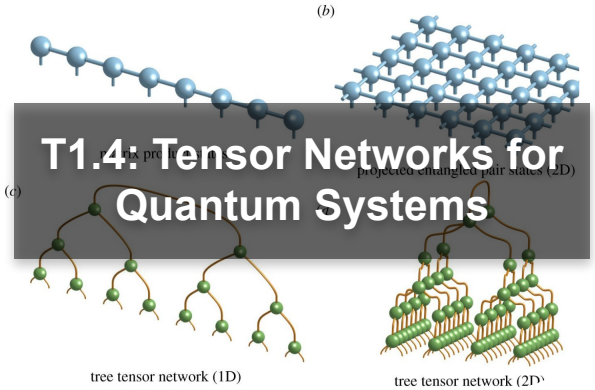
WP4: Education Programmes and Outreach

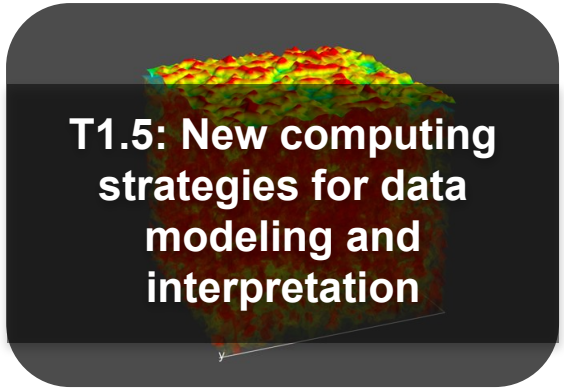
What follows is a personal selection of highlights. For details on all the amazing work, check out the 2nd NGT technical workshop

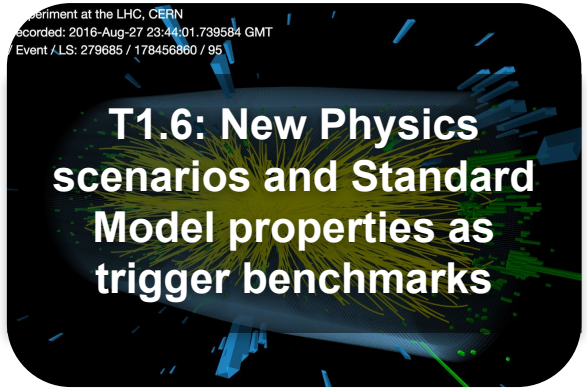
WP1: Infrastructure, Algorithms and Theory

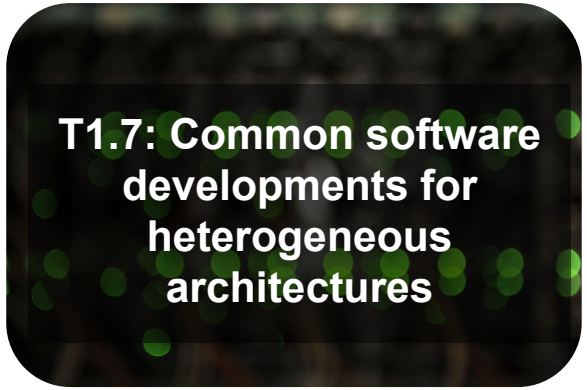












1.1 Hardware & services for large scale NN optimisation etc.

Focus: designing, procuring, deploying and operating the computing infrastructure (hardware and software) and platforms required to support

- **common tasks in WP1**
  - Hardware-aware neural network training workflows
  - Next-generation physics simulations
- **specific activities in WP2 and WP3**

- **Resources procured and installed in the new CERN Data Center**
- **Specialized CPU nodes eg for firmware synthesis**

Nvidia AMD

Description	Nodes	Cores	RAM	GPU	Network	nvidia.com/gpu.product
Nvidia H100 188GB NVL	12	192	1.5TB	8	200G	NVIDIA-H100-NVL
Nvidia H100 80GB SXM	6	64	1TB	4	400G	NVIDIA-H100-80GB-HBM3
Nvidia L40S 48GB	7	128	1.5TB	4	200G	NVIDIA-L40S

Nvidia AMD

Description	Nodes	Cores	RAM	GPU	Network	amd.com/gpu.product-name
AMD Instinct MI300X 192GB	2	128	1.5TB	8	200G	AMD_Instinct_MI300X_OAM
AMD Radeon Pro W7900 48GB	6	128	770GB	4	200G	AMD_Radeon_PRO_W7900

1.1 Hardware & services for large scale NN optimisation etc.

Focus: designing, procuring, deploying and operating the computing infrastructure (hardware and software) and platforms required to support

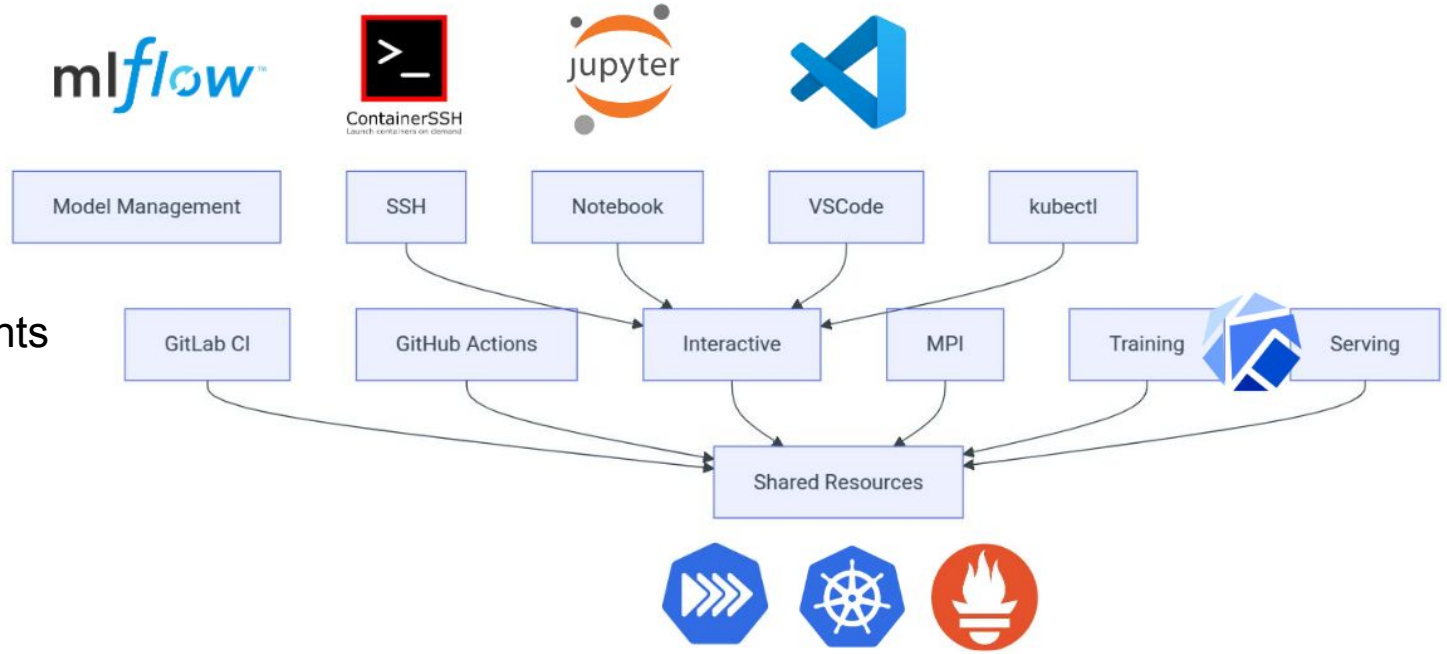
- **common tasks in WP1**
  - Hardware-aware neural network training workflows
  - Next-generation physics simulations
- **specific activities in WP2 and WP3**

- **Resources procured and installed in the new CERN Data Center**
- **Specialized CPU nodes eg for firmware synthesis**

1.1 Hardware & services for large scale NN optimisation etc.

Service platform based on kubernetes

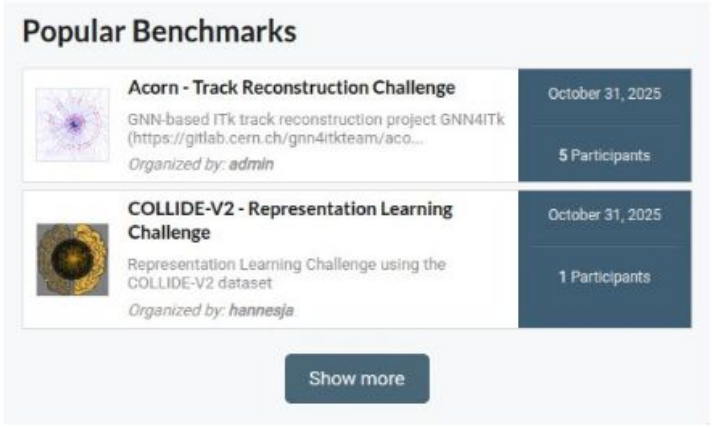
- **Monitoring set up, including utilization, power dissipation, …**
- **Model management using mlflow**
  - Track and compare multiple model experiments
  - Built-in tools and charts to evaluate and visualize model performance
  - Storage of model artifacts, plots, …
  - Share experiments and results across teams



1.1 Hardware & services for large scale NN optimisation etc.

Platform to host Machine Learning Challenges

- **On-premises deployment similar to Kaggle**
  - Leaderboard with scoring and ranking
- **Maintainers create competitions and define scoring metrics**
- **Users submit code and/or predictions that are evaluated against a test set**



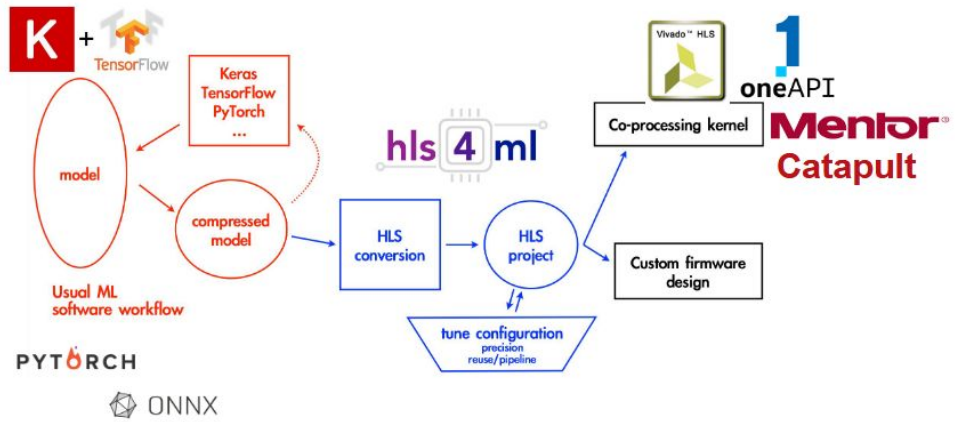
1.2 Fast Inference of complex network architectures…
1.3 Hardware-aware AI optimization

Focus of 1.2:

- **Develop hardware-aware inference engine + advanced code-generation core**
- **Utilize and expand hls4ml: source-to-source compiler creating hw-optimal HLS designs for FPGAs**
  - New model architectures
  - Support more deployment options
- **Ensure the work benefits the experiment-specific work packages, but also the FastML/hls4ml communities**

Focus of 1.3:

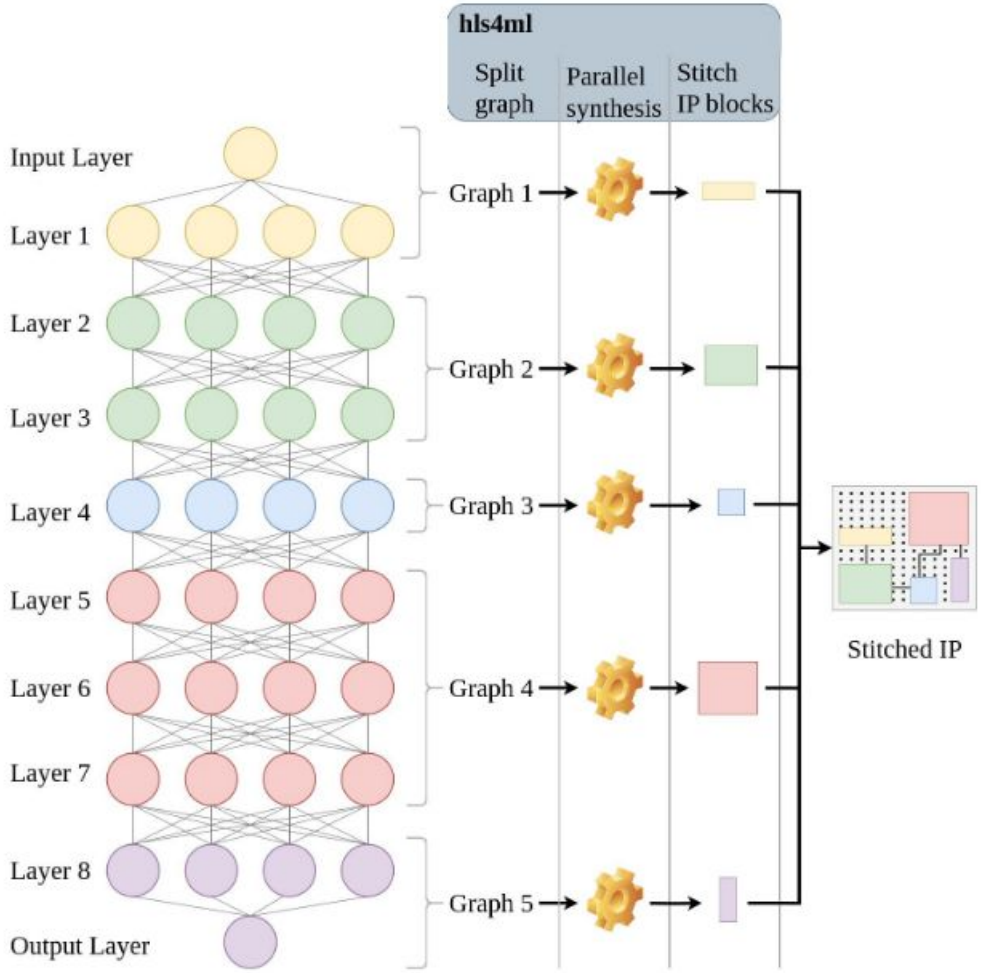
- **Building a library of model compression techniques with a unified interface**
- **Collection of proven methods and set of recommendations for end-users not requiring extensive knowledge of efficient ML and HW specifics**



1.2 Fast Inference of complex network architectures…

Model Graph Splitting for Parallel Synthesis

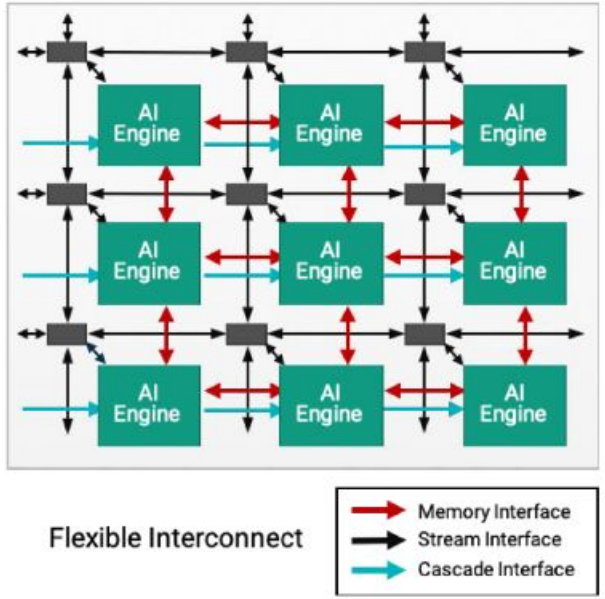
- **Partition model graph in hls4ml into smaller, independent subgraphs to reduce synthesis times (complex models often > 24h!)**
- **Workflow:**
  - Select layers as splitting points, currently defined user
  - Parallel synthesis, exporting each subgraph as IP block
  - Automatic merging in Vivado to stitch final IP



1.2 Fast Inference of complex network architectures…

AMD AI Engines support

- **Exploring AI engines for more flexible and future ready triggers**
  - Man offer higher compute density and on-chip memory than programmable logic on FPGA
  - May handle better medium to large neural networks
  - Short compilation times
  - Co-existence with FPGA fabric
  - However:
    - Generic and scalable designs difficult
    - Limited toolchain support, low-level development
- **Development of aie4ml**
  - Companion package to hls4ml
  - Select ‘AIE’ as backed
  - Scales workloads across AIE array automatically
  - Supports Gen2/Gen3 devices (broader support planned)



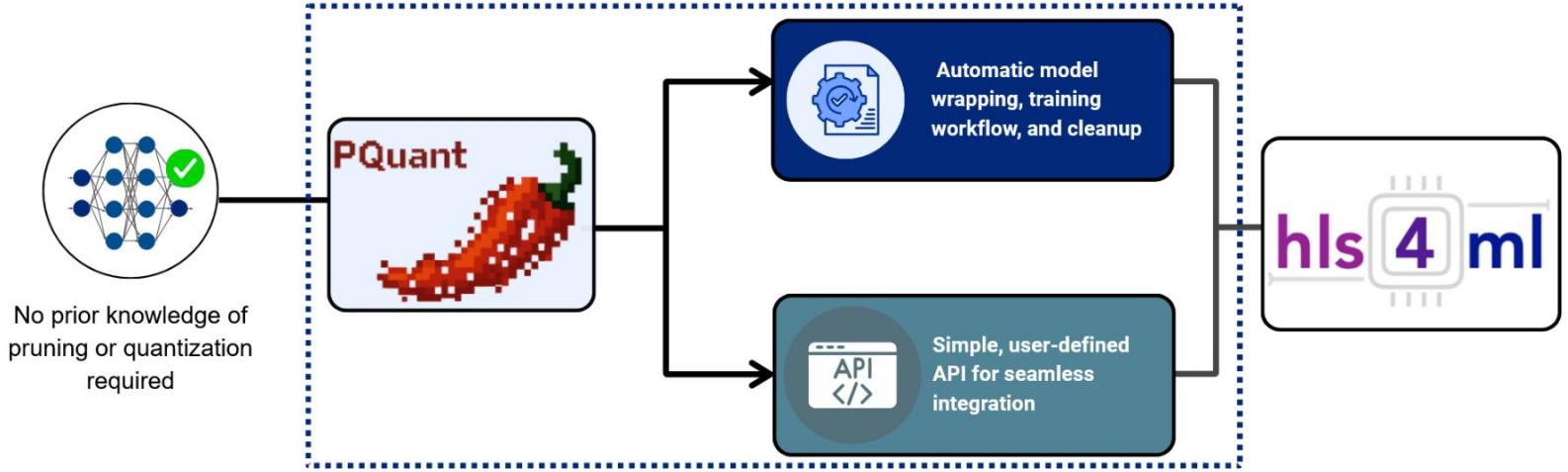
aie4ml

arxiv:2512.15946

1.2 Fast Inference of complex network architectures…
1.3 Hardware-aware AI optimization

Integration of PQuantML with hls4ml

- **In 1.3 development of PQuantML to replace TF/QKeras in the compression workflow → see presentation by Roope**
- **Direction integration with hls4ml:**
  - Directly parse the model with PQuantML layers
  - No need to manually tweak the config dictionary
  - Assured bit-exactness thanks to the *bit_exact* optimizer pass



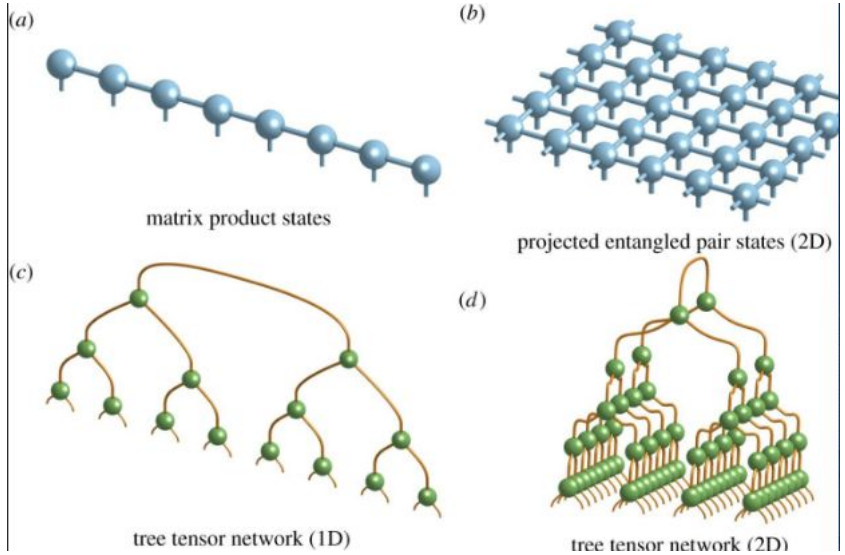
1.4 Tensor Networks for Quantum Systems

- **Consider quantum state of N interacting particles (eg qubits)**
  - Each particle has d possible states
    → d^N total states!
  - Example: 50 qubits → 10^{50} numbers!
  - Storing or manipulating impossible
- **Matrix Product State**

$$|\Psi\rangle = \sum_{\{s_j\}=1}^{d} Tr\{A[s_1]\cdots A[s_N]\}|s_1\rangle\cdots|s_N\rangle$$

reduces parameter count and computation cost

$$|\psi\rangle = \sum_{i_1,i_2,\cdots,i_N=1}^{d} \psi_{i_1,i_2,\cdots,i_N}|i_1, i_2, \cdots, i_N\rangle$$



1.4 Tensor Networks for Quantum Systems

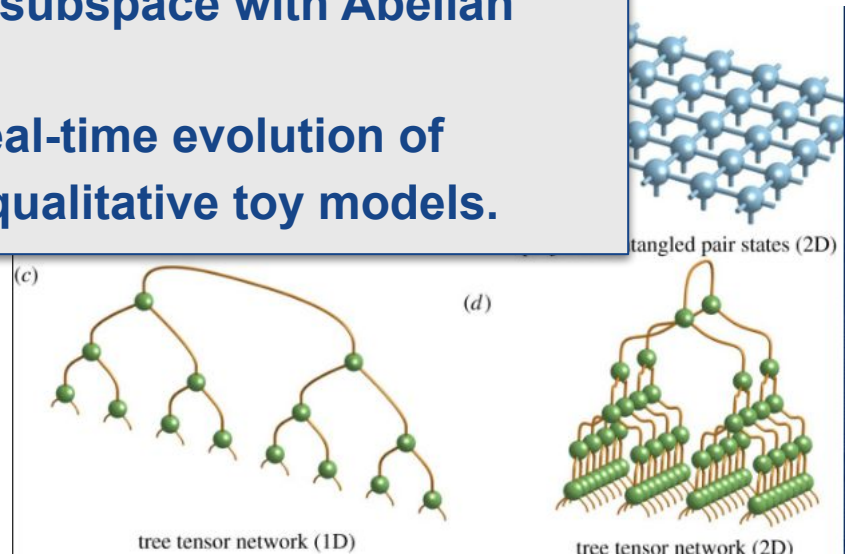
- **Consider quantum state of N interacting particles (eg qubits)**
  - Each partic
    → d^N tota
  - Example: 5
  - Storing or r
- **Matrix Produc**

$$|\psi\rangle = \sum_{i_1,i_2,\cdots,i_N=1}^{d} \psi_{i_1,i_2,\cdots,i_N} |i_1, i_2, \cdots, i_N\rangle$$

$$|\Psi\rangle = \sum_{\{s_j\}=1}^{d} Tr\{A[s_1]\cdots A[s_N]\}|s_1\rangle\cdots|s_N\rangle$$

reduces parameter count and computation cost

- **MPS and TNS in higher dimensions are exact descriptions of the "physical" gauge invariant subspace with Abelian and non-Abelian symmetry**
- **TNS allow the simulation of the real-time evolution of Lattice Gauge Theory, at least in qualitative toy models.**



1.4 Tensor Networks for Quantum Systems

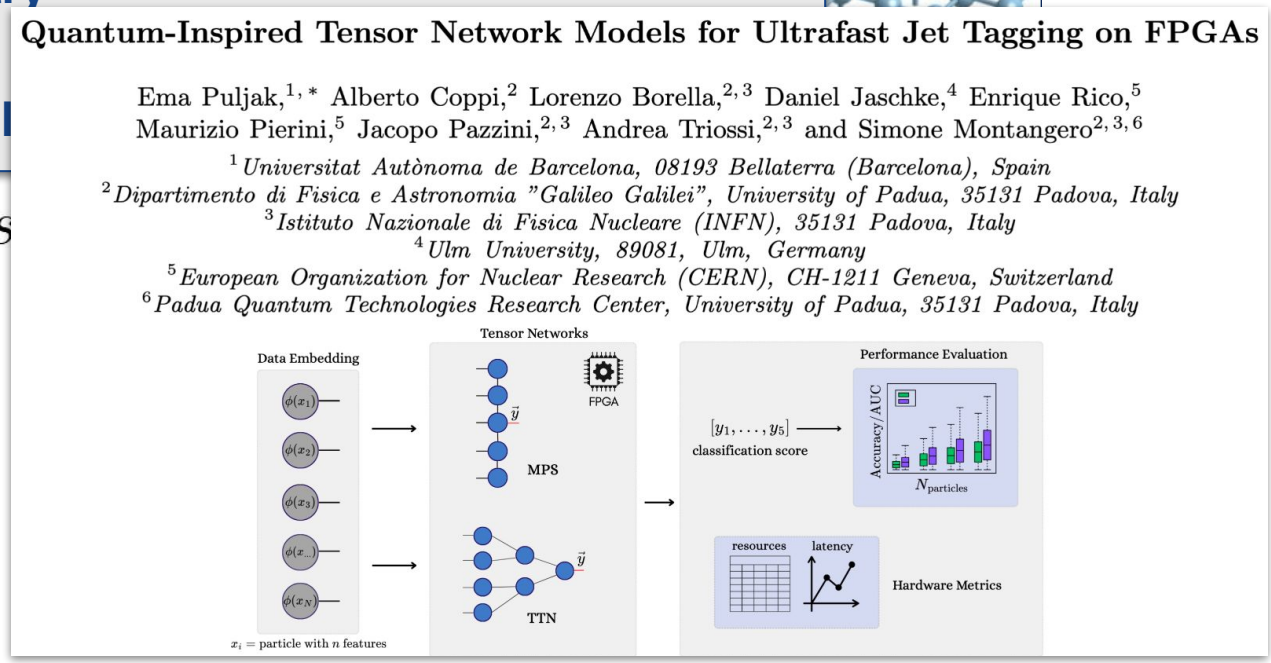
- **Consider quantum state of N interacting particles (eg qubits)**
  - Each partic
    → d^N tota
  - Example: 5
  - Storing or r
- **Matrix Produc**

$$|\psi\rangle = \sum_{i_1,i_2,\cdots,i_N=1}^{d} \psi_{i_1,i_2,\cdots,i_N} |i_1, i_2, \cdots, i_N\rangle$$

$$|\Psi\rangle = \sum_{\{s_j\}=1}^{d} Tr\{A[s_1]\cdots A[s_N]\}|s_1\rangle\cdots|s$$

reduces parameter count and computation cost

- **MPS and TNS in higher dimensions are exact descriptions of the "physical" gauge invariant subspace with Abelian and non-Abelian symmetry**
- **TNS allow the simulation**
  Lattice Gauge Theory, at l



Quantum-Inspired Tensor Network Models for Ultrafast Jet Tagging on FPGAs

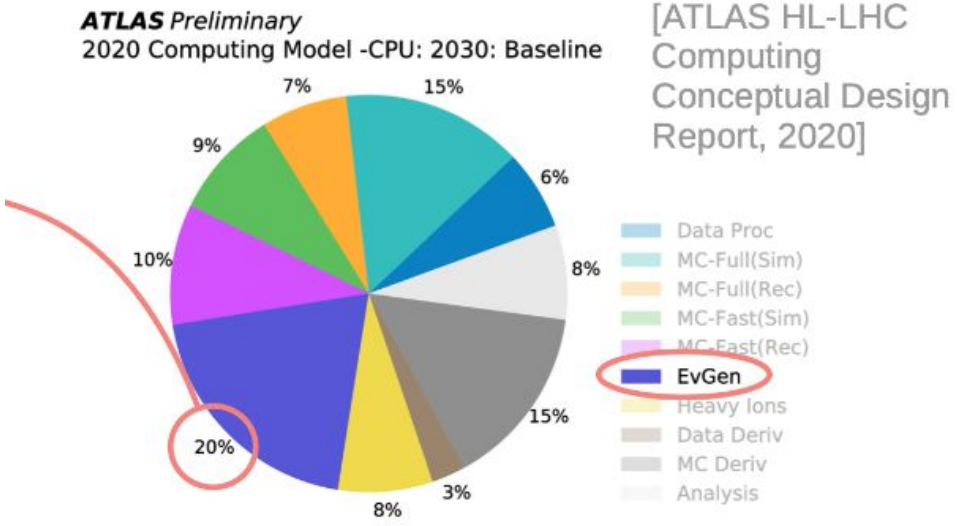
Ema Puljak,[1,*] Alberto Coppi,[2] Lorenzo Borella,[2,3] Daniel Jaschke,[4] Enrique Rico,[5]
Maurizio Pierini,[5] Jacopo Pazzini,[2,3] Andrea Triossi,[2,3] and Simone Montangero[2,3,6]

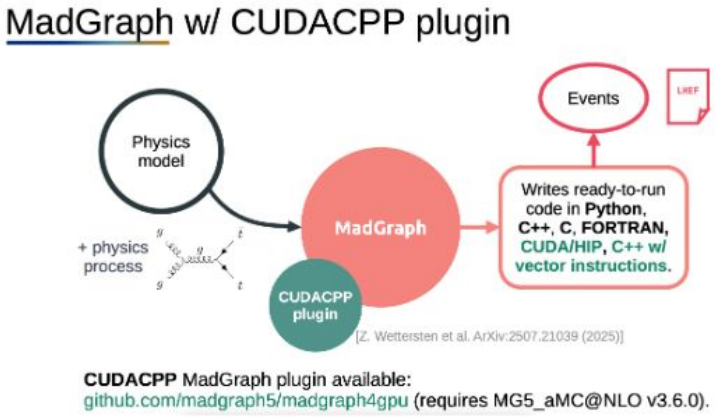
[1] *Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain*
[2] *Dipartimento di Fisica e Astronomia "Galileo Galilei", University of Padua, 35131 Padova, Italy*
[3] *Istituto Nazionale di Fisica Nucleare (INFN), 35131 Padova, Italy*
[4] *Ulm University, 89081, Ulm, Germany*
[5] *European Organization for Nuclear Research (CERN), CH-1211 Geneva, Switzerland*
[6] *Padua Quantum Technologies Research Center, University of Padua, 35131 Padova, Italy*

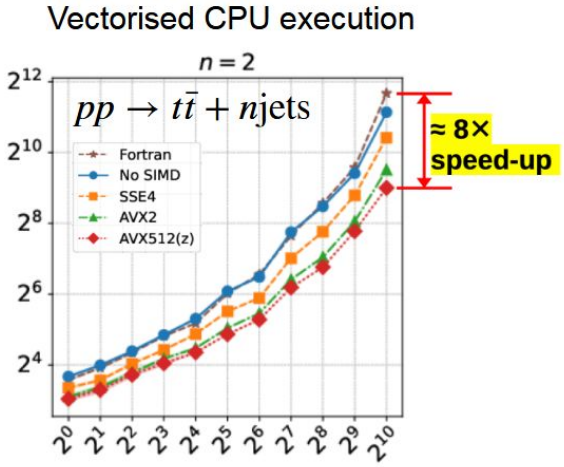
1.5 New computing strategies for data modelling …

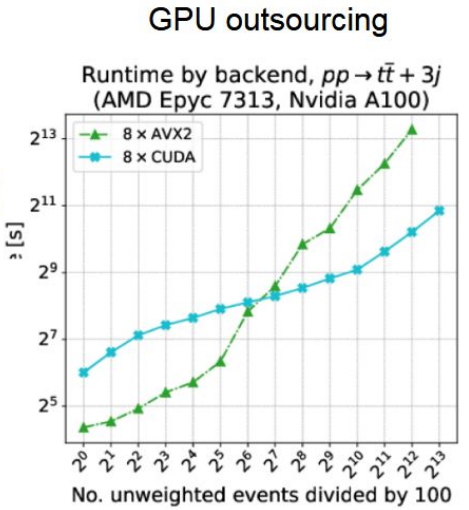
Monte-Carlo event generation for HL-LHC estimated to be ≈ 20 % of total computing resources

- **Focus:**
  - Evolve current event generation strategies and codes, porting and optimizing to new architectures (CPU, GPU), exploiting ML strategies
  - Acceleration of matrix element evaluation
  - Prepare MC generator framework for higher-order calculations





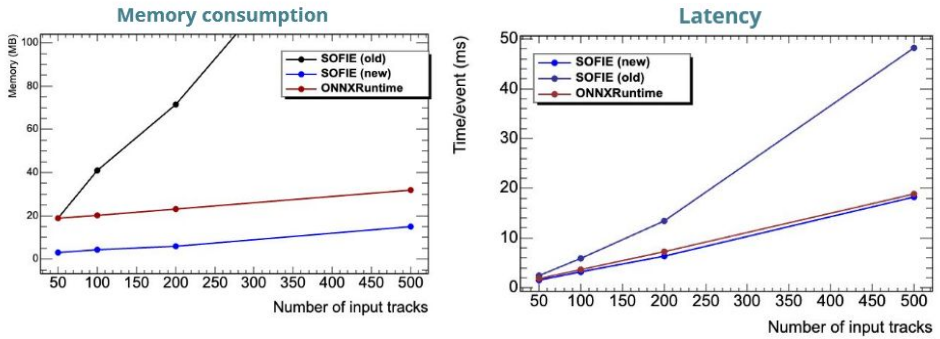




1.7 Common software for heterogeneous architectures

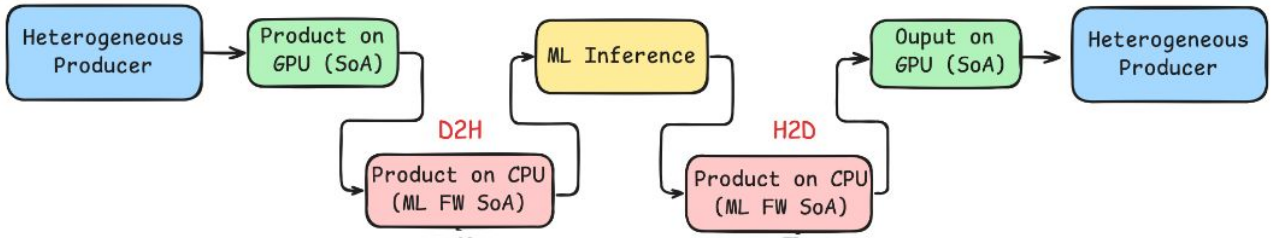
Inference with ROOT/TMVA SOFIE (repository)
(System for Optimized Fast Inference code Emit)

- **Reads ML models**
- **Generates C++ code for inference**
- **Currently eg optimizing memory and latency on CPU**



Direct Inference with PyTorch

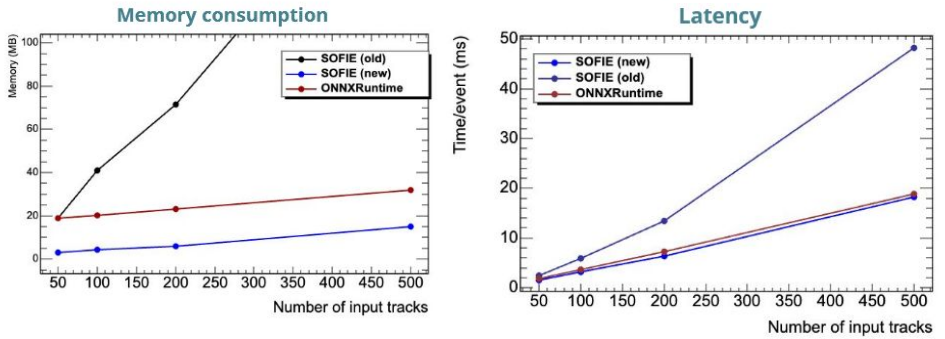
- **Integration of PyTorch in alpaka**
- **Avoiding device-host-device copies by conversion of SoA data directly to PyTorch tensors**



1.7 Common software for heterogeneous architectures

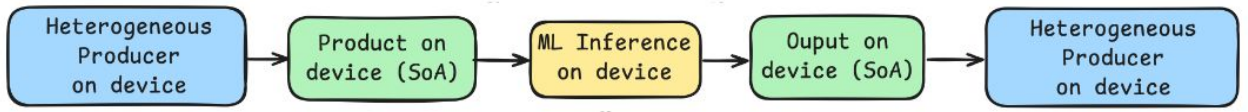
Inference with ROOT/TMVA SOFIE (repository)
(System for Optimized Fast Inference code Emit)

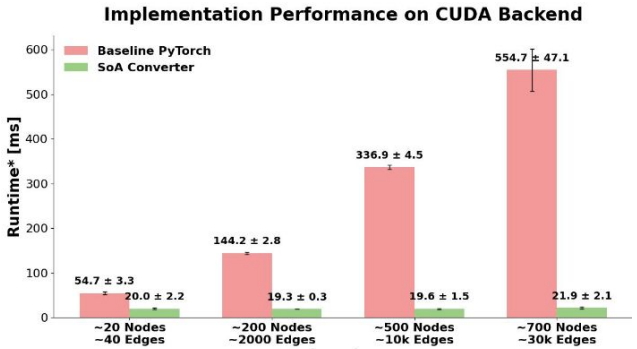
- **Reads ML models**
- **Generates C++ code for inference**
- **Currently eg optimizing memory and latency on CPU**



Direct Inference with PyTorch

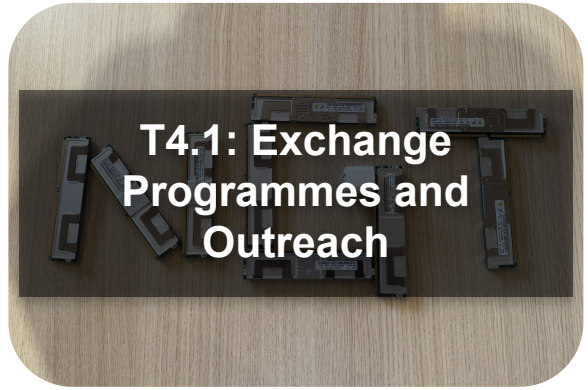
- **Integration of PyTorch in alpaka**
- **Avoiding device-host-device copies by conversion of SoA data directly to PyTorch tensors**

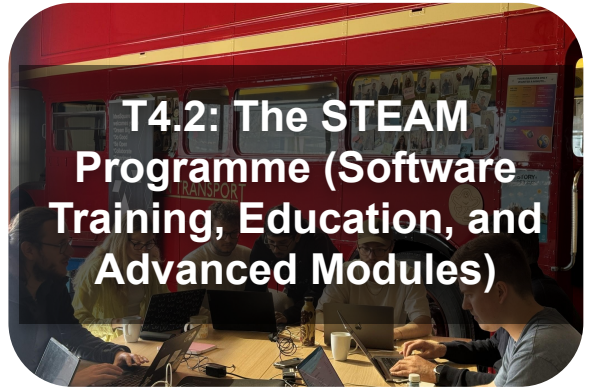




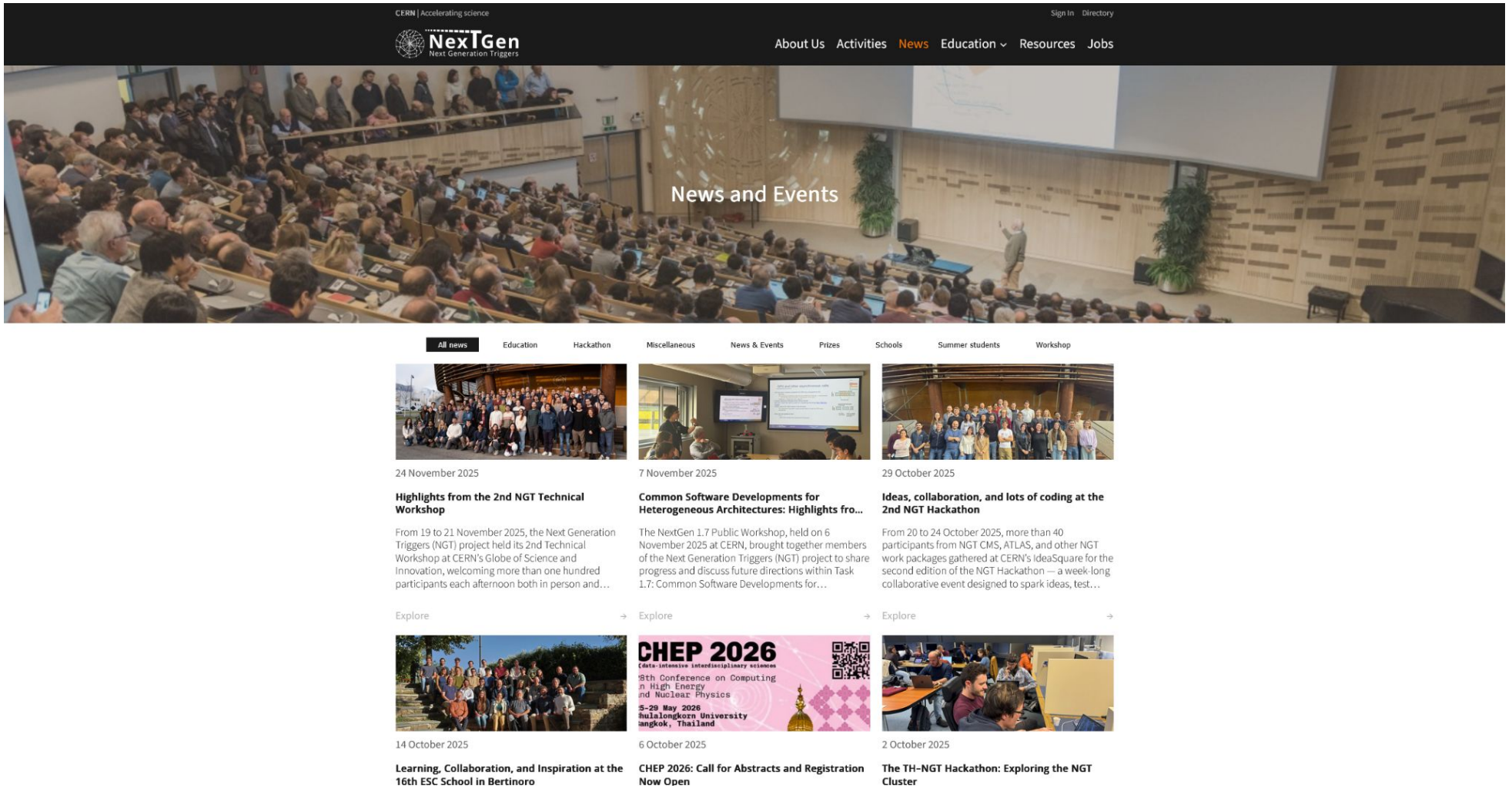
WP4: Education Programmes and Outreach

- **Train, sustainably**
  - Develop project members knowledge with an extensive education programme and multiple training opportunities
  - Reuse and enhance existing infrastructures to cover education and training needs of NextGen staff
- **Exchange, impact beyond NextGen**
  - Visiting scientists, public seminars
  - Strong focus on outreach and external communication
  - Make NextGen knowledge accessible also outside the project
    - Conference participation and online presence
    - Support NextGen staff to contribute to Outreach initiatives, Schools, and Education activities

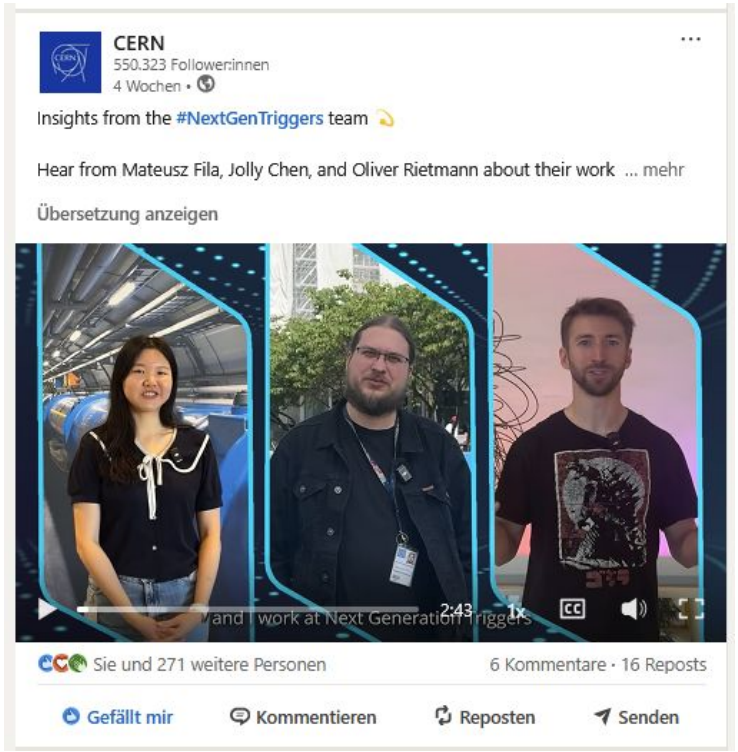


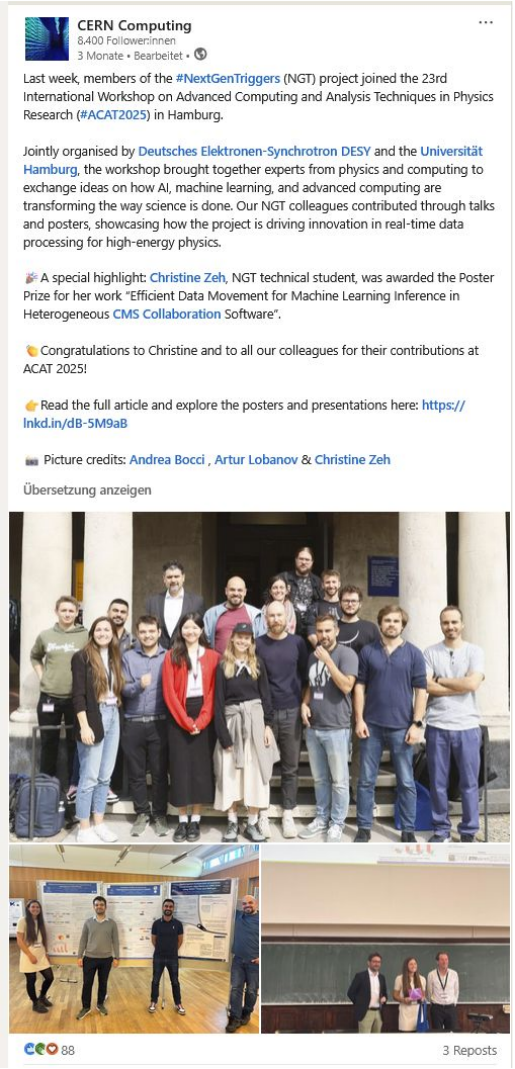


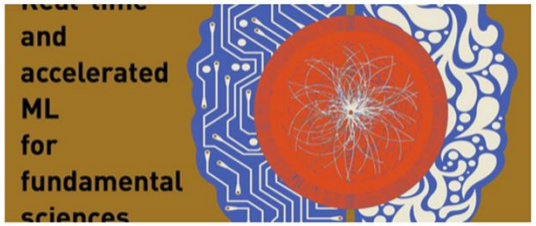
Website https://nextgentriggers.web.cern.ch/



Outreach and events







27 February 2025

Fast Machine Learning for Science Conference 2025

The Fast Machine Learning for Science Conference 2025 will take place from September 1–5, 20 ETH Zurich, bringing together researchers ar industry experts to explore the latest advanc in machine learning (ML) for scientific discov

10 February 2025

ISOTDAQ 2025 – International School of Trigge and Data Acquisition

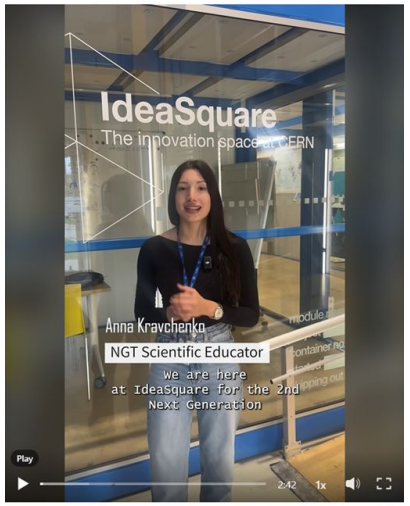
ISOTDAQ 2025 is back for its 15th edition, offering a unique opportunity for students and early-career researchers in physics, engineering, and computing

Recap of the 16th ESC School in Bertinoro

Participants of the ESC2025, in Bertinoro, Italy. Picture credit: Andrea Bocci.

As part of our ongoing commitment to training and development, the Next Generation Triggers project continues to support its members in strengthening their expertise in computing and data science. This year, several of our colleagues travelled to Bertinoro, Italy, to take part in the Efficient Scientific Computing (ESC) School 2025, held from 29 September to 9 October 2025.

People interviewed:
Giovanni Zago Wahid Redjeb Kyungmin Park Elisa Fontanesi Maciej Glowacki Anna Kravchenko Dimitrios Danopoulos



Wahid Redjeb and 79 others
7 reposts

2nd NGT Hackathon @ CERN Idea Square

Importance of Education in NextGen

- **HL-LHC Trigger/DAQ/Software challenges demand specialists**
- **Problem: Knowledge draining in the time scale of HL-LHC**
- **How do we retain knowledge and build a core community of experts that can last long term? How do we make a strong impact onto HEP scientific community?**
  - Work in synergy with existing schools and tutorials avoiding overlaps and building on top of these strong foundations.
  - Creating opportunities for experts and young scientists to work together.
  - Collaborate with existing experts and universities to create a portfolio of intensive courses and tutorials.

Schools supported in 2025

School	Dates	Registration Deadline	Location	Link
CERN School of Computing on Machine Learning 2025	8–14 June 2025	12 March 2025	Malmö, Sweden	Visit page
ISOTDAQ School	17–26 June 2025	14 February 2025	Vilnius, Lithuania	Visit page
CERN School of Computing 2025	6–19 July 2025	9 April 2025	Lund, Sweden	Visit page
MLx Representation Learning & Gen.AI	7–10 August 2025	14 June 2025	Oxford, UK / Online	Visit page
INFN ESC – Efficient Scientific Computing School	28 Sept – 9 Oct 2025	30 June 2025	Italy	Visit page
tCSC on Heterogeneous Computing	5-11 Oct 2025	10 July 2025	Split, Croatia	Visit page
HEP C++ Course and Hands-on Training (Advanced C++)	27-31 Oct 2025	—	CERN	Visit Page
1st CERN School of Computing in Latin America	11-24th January 2026	1 October 2025	Santiago, Chile	Visit page

NGT Learning Hub https://nextgentriggers.web.cern.ch/our-learning-hub/

Alpaka Tutorial

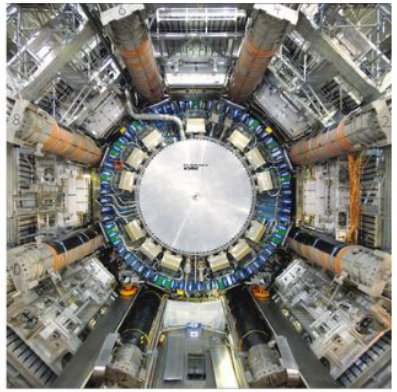
By Andrea Bocci

CMSSW Tutorial

By Felice Pantaleo

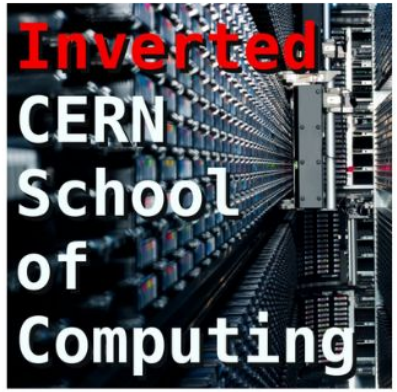
Git Tutorial

By Simone Rossi Tisbeni



ATLAS Tutorials

By ATLAS Community

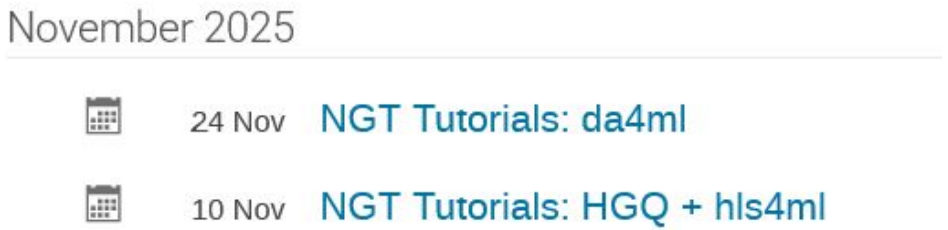


Inverted CSC

By IT Community

Tutorials

- Series of tutorials started on hls4ml, HGQ, da4ml, and the tools under development in Tasks 1.2 and 1.3
- Open also to anyone outside of NGT!
- Materials (and recordings) available here https://indico.cern.ch/category/20359/
- Tutorials on pQuantML and aie4ml coming up in the next months!



CERN STEAM Academy

Based at CERN, Geneva

22 June - 28 August 2026

A 10-weeks long summer advanced training programme at CERN combining Software Technologies, Edge Computing, Data Science and ML.

A curriculum for future Data acquisition, Edge computing, Advanced software and Data science experts.

Audience:

- PhD candidates and postdoctoral researchers;
- Researchers or engineers with a completed Master's degree.

https://nextgentriggers.web.cern.ch/cern-steam-academy/ **Applications open until Februrary 16th!**

more info in the backup!

Programme Themes & Key Topics

Theme 1

Software Technologies

- Advanced C++
- Algorithms
- Floating-point
- Efficient Python
- Architecture & memory model
- Parallel & heterogeneous computing
- Developer tools & workflows

Theme coordinators - Noemi Calace, Vincenzo Innocente, David Rohr, Giulio Eulisse.

Theme 2

Edge Computing

- FPGAs with VHDL
- High-Level Synthesis
- AI at the edge with hls4ml and FINN
- AI Engines
- SoC FPGAs
- Networking
- High-speed optical links
- Storage

Theme coordinator - Hannes Sakulin, Barthelemy Von Haller, Wainer Vandelli, Paolo Durante, Tommaso Colombo.

Theme 3

Data Science & ML

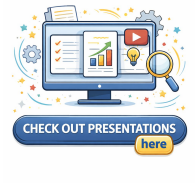
- MLOps
- RNNs, GNNs, transformers
- Autoencoders & anomaly detection
- Generative models
- Foundation models
- LLMs/RAGs/agents;
- Reinforcement learning
- Interpretability & uncertainty
- Pruning & quantisation

Theme coordinator - Lorenzo Moneta.

Summary

NGT is in full swing, with the ultimate goal to expand event selection acceptance of ATLAS and CMS @ HL-LHC

NGT farm running, ML challenge platform, expansion of FastML tools, education programme, CERN STEAM academy



More details on the technical work, also on tasks not covered in this talk, check out the presentations @ 2nd technical workshop


CERN
NexTGen
Next Generation Triggers

Backup

What the Academy Provides

For students

- Accommodation and meals for external students for the full 10 weeks
- Scheduled group visits to CERN facilities and experiments.
- Social activities and local excursions alongside the academic programme.
- A CERN access card.
- A CERN computing account.
- Official invitation letters for Visa (if needed).

https://nextgentriggers.web.cern.ch/cern-steam-academy/ **Applications open until Februrary 16th!**

Programme Structure & Participation Model