



Novel Ideas for Next-Generation Triggers (NGT) and DAQ at the HL-LHC

Mateusz Zarucki (CERN)
*on behalf of the
ATLAS and CMS Collaborations*

mateusz.zarucki@cern.ch

Corfu2025 Workshop
on Future Accelerators

1st May, 2025



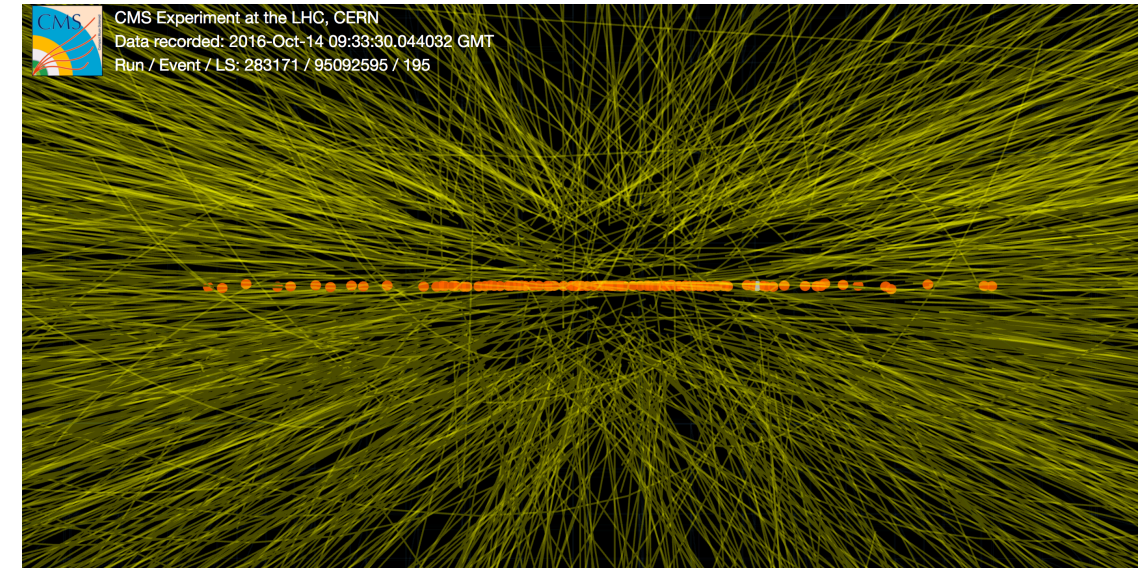
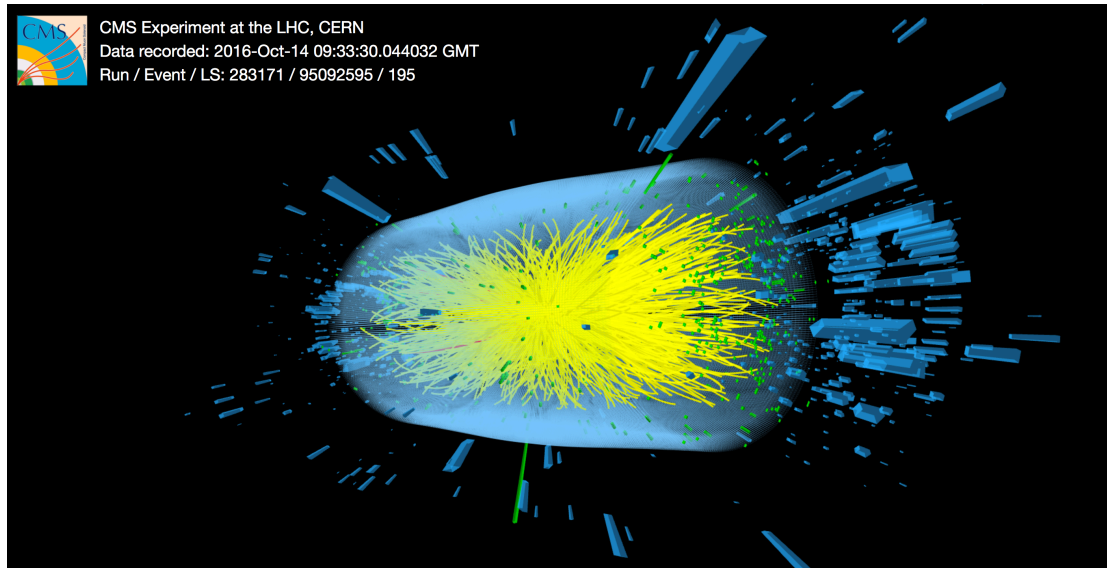
NextGen
Next Generation Triggers



NextGen

High-Luminosity LHC (HL-LHC)

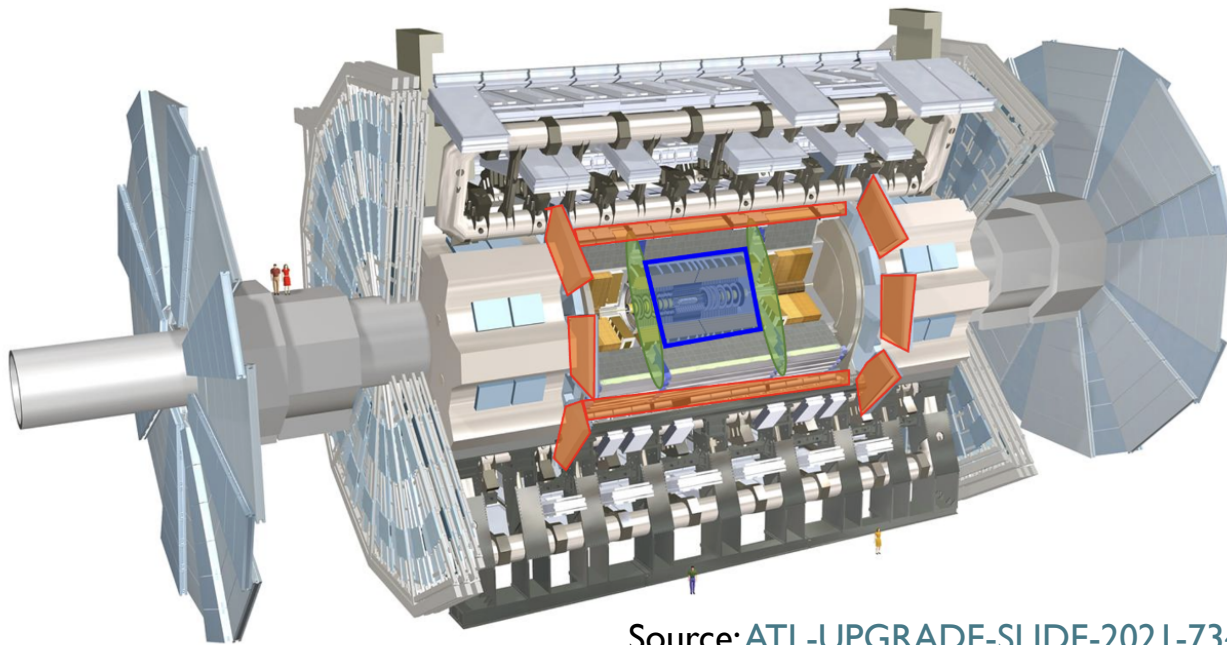
- Large Hadron Collider (LHC) will undergo a major upgrade to the High-Luminosity LHC (HL-LHC) by 2030
 - delivering 10x more data ($\approx 3000 \text{ fb}^{-1}$) between 2030 - 2040 than until now, to probe rare phenomena
- 3-4x higher instantaneous luminosity & pileup interactions: $L \approx 5 - 7.5 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$, pileup $\langle \text{PU} \rangle \approx 140\text{-}200$



- ATLAS and CMS are preparing major Phase-2 upgrades in all sub-systems to cope with these conditions
 - including the trigger and DAQ systems with L1/L0 rate $\approx 500 \text{ kHz} - 1 \text{ MHz}$, HLT/EFV rate $\approx 10 \text{ kHz} + \text{scouting/TLA}$

ATLAS Phase-2 Upgrades

ATLAS is preparing major Phase-2 upgrades in all systems to cope with these conditions, encompassing multiple detectors and systems, including tracking, triggering and data acquisition:



Source: [ATL-UPGRADE-SLIDE-2021-734](#)

Upgraded Trigger and Data Acquisition System

- Single Level Trigger with 1 MHz output
- Improved 10 kHz Event Filter (EF)

Electronics Upgrades

- On-detector/off-detector electronics upgrades of LAr Calorimeter, Tile Calorimeter & Muon Detectors
- 40 MHz continuous readout with finer segmentation to trigger

High Granularity Timing Detector (HGTD)

- Precision time reconstruction (30 ps) with Low-Gain Avalanche Detectors (LGAD)
- Improved pile-up separation and bunch-by-bunch luminosity

New Muon Chambers

- Inner barrel region with new RPCs, sMDTs, and TGCs
- Improved trigger efficiency/momentum resolution, reduced fake rate

New Inner Tracking Detector (ITk)

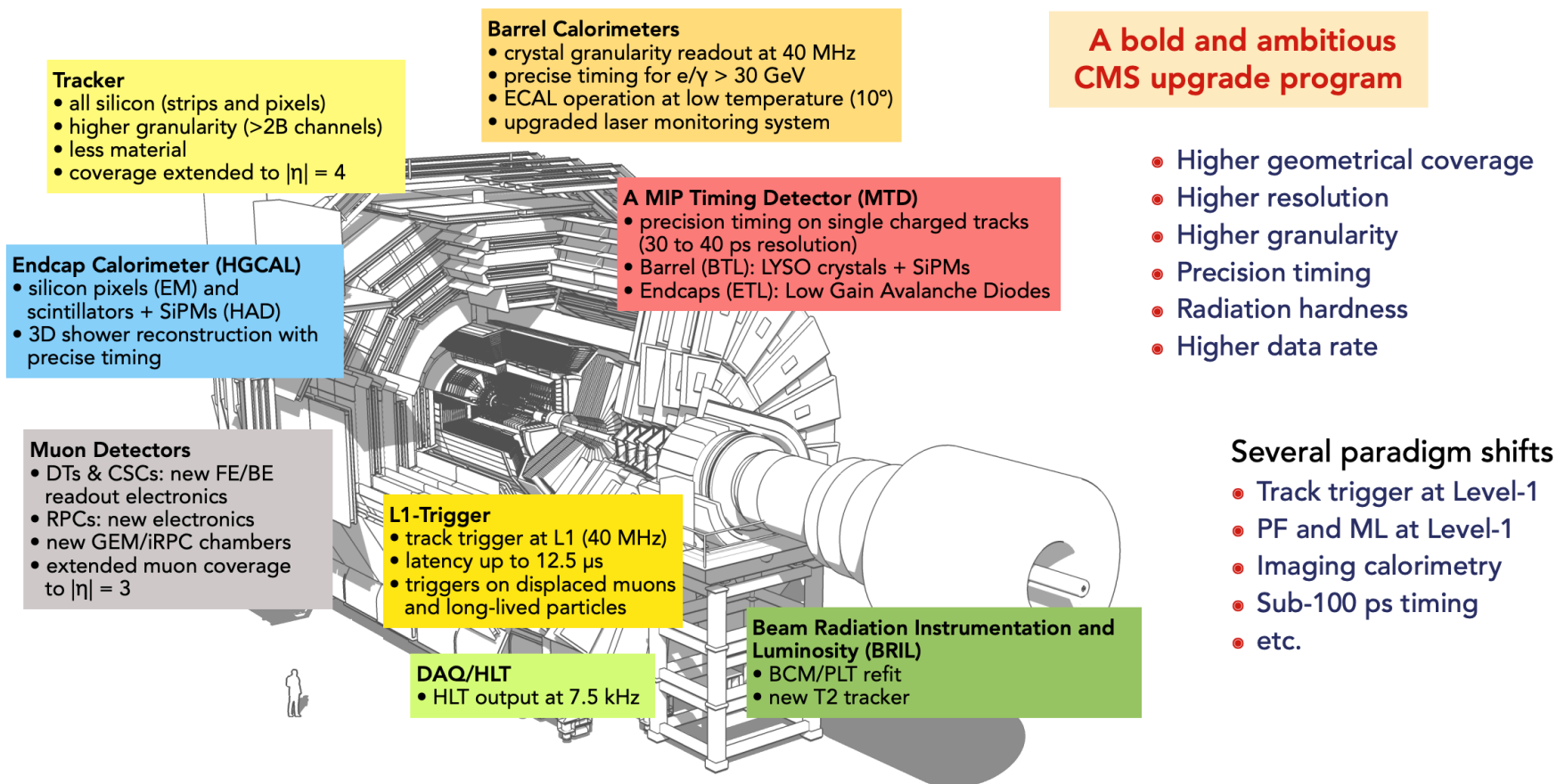
- All silicon with 5 pixel layers, 4 double-sided strips in the barrel, $|\eta| < 4$ with at least 9 hits
- Less material, finer segmentation

Additional small upgrades

- Luminosity detectors (1% precision)
- HL-ZDC (Heavy Ion physics)

CMS Phase-2 Upgrades

CMS is preparing major Phase-2 upgrades in all systems to cope with these conditions, encompassing multiple detectors and systems, including tracking, triggering and data acquisition:



Trigger (Run 3)

"the trigger does not decide which physics model is right, it just decides which physics model is left"

It is not feasible to readout and record every event (≈ 30 MHz) due to hardware limitations

- nor is it always efficient in terms of the physics programme of the experiment ('contaminated' by less-interesting inelastic and QCD-multijet events)

DAQ and trigger systems are designed to analyse, filter and collect the collision data at these enormous rates to select only the most interesting events for offline analysis → encapsulating the scientific programmes of ATLAS & CMS

- e.g. isolated/non-isolated muons/electrons, photons, taus, jets, missing transverse momentum

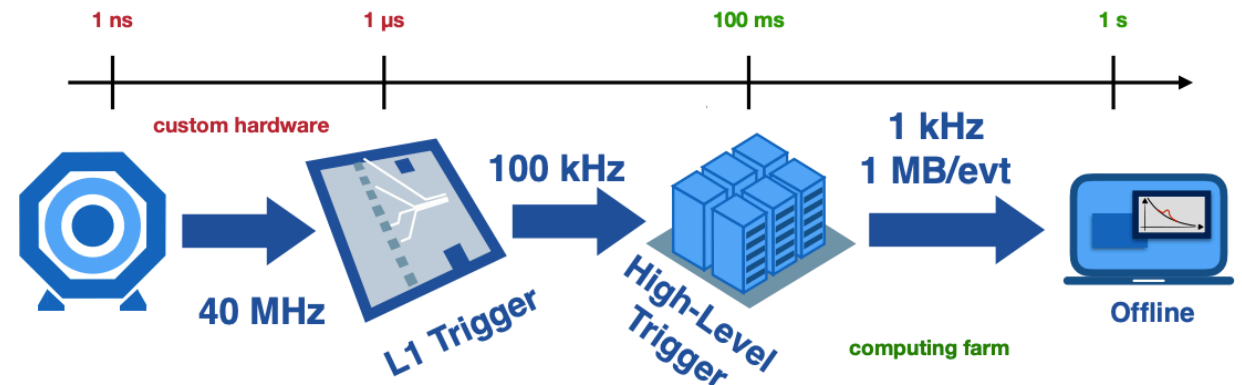
Run 3 conditions: instantaneous luminosity $L \approx 2 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ and pileup $\langle \text{PU} \rangle \approx 65$

- CMS/ATLAS Level-I Trigger (L1T) based on hardware reduces the event rate to ≈ 100 kHz
- CMS High Level Trigger (HLT)/ATLAS Event Filter (EF) based on software down to ≈ 3 kHz
 - [+ 3.5 kHz parking + 8 - 20 kHz scouting]

bunch crossing rate (25 ns):

$$f_{\text{BX}} = 40 \text{ MHz}$$

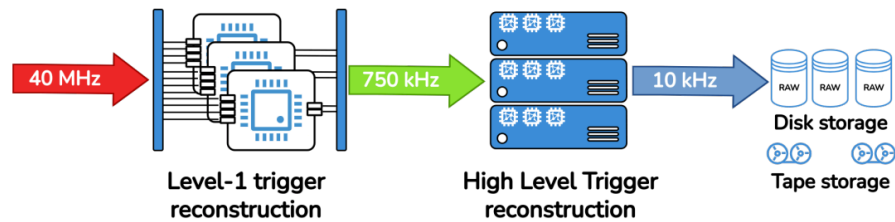
(effective ≈ 30 MHz)



Trigger (HL-LHC and Phase-2)

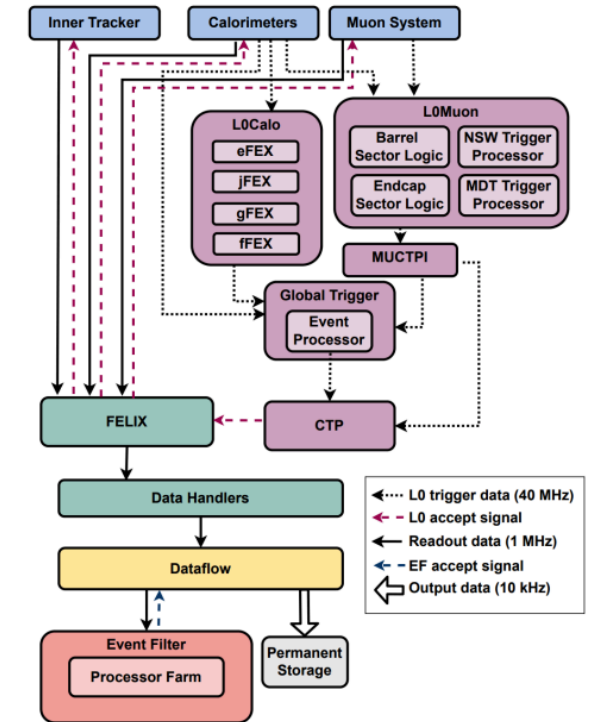
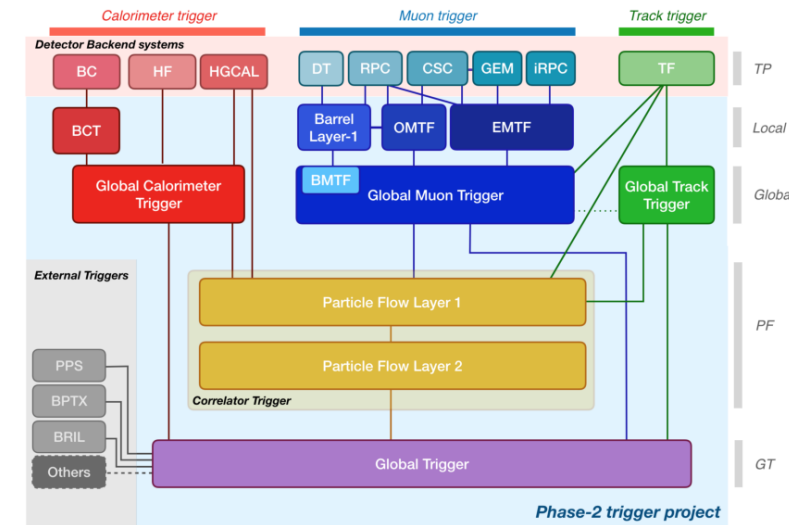
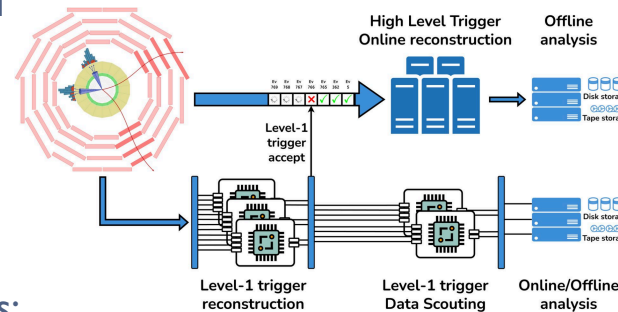
Baseline **CMS** trigger and DAQ design for the HL-LHC foresees:

- **Level-1 trigger (L1T)**, implemented in custom hardware, analysing events at 40 MHz and selecting up to 750 kHz, based on calorimeter and muon system data
 - additionally including track reconstruction and particle flow (PF) capabilities
 - L1T scouting system, analysing all events at 40 MHz to extract physics signatures on the fly
- **High Level Trigger (HLT)**, implemented in software on a computing farm for high-level event reconstruction and selection, analysing L1-accepted events at 750 kHz for a further rate reduction selecting up to 10 kHz for offline processing and long-term storage
 - includes a dedicated HLT scouting programme in parallel



Baseline **ATLAS** trigger and DAQ design for HL-LHC foresees:

- **Level-0 (L0) trigger** implemented in custom hardware based on calorimeter and muon system data, with a maximum nominal rate of selected events of 1 MHz
- **Event Filter (EF)**, trigger implemented in software on a computing farm for high-level event reconstruction and selection, analysing L0-accepted events, adding tracker information for a further rate reduction, up to 10 kHz for offline processing and long-term storage
 - includes a dedicated trigger-level analysis (TLA) programme in parallel



Next-Generation Triggers (NGT)

Next-Generation Triggers (NGT) is a cross-disciplinary, cross-experiment project to leverage innovative computing technologies for data acquisition and processing, broken down into four work packages:



Next-Generation Triggers (NGT)

Next-Generation Triggers (NGT) is a cross-disciplinary, cross-experiment project to leverage innovative computing technologies for data acquisition and processing, aiming to go beyond the baseline Phase-2 upgrades

- **WP1: Infrastructure, Algorithms and Theory**

- 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**

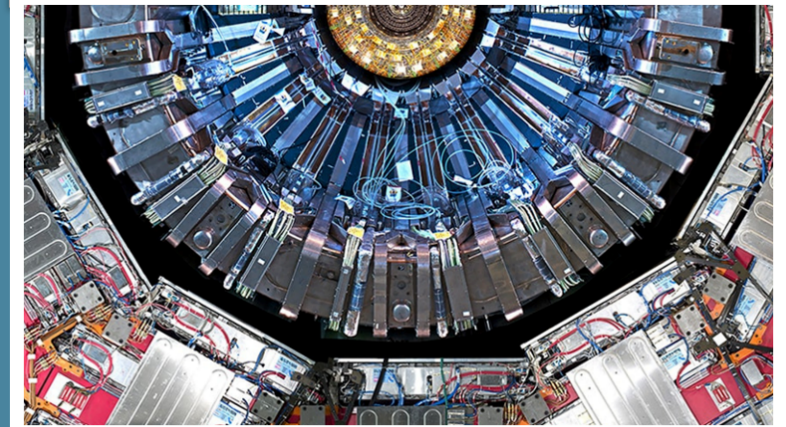
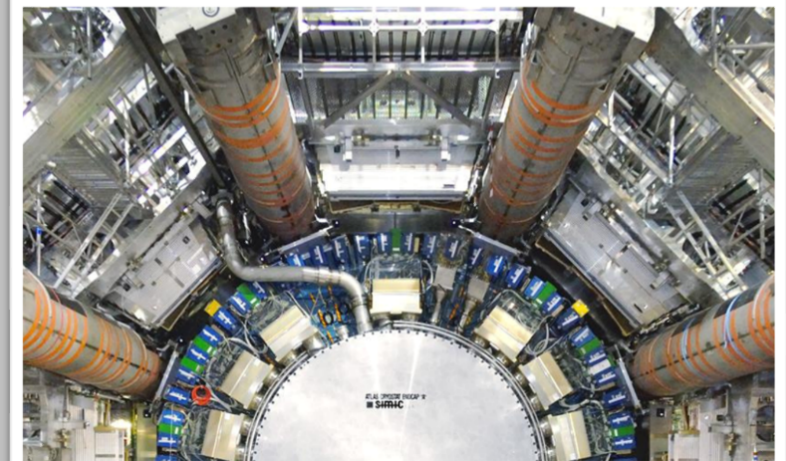
- focus on improved and accelerated filtering and exotic signature detection

- **WP3: Rethinking the CMS Real-Time Data Processing**

- design a novel AI-powered real-time processing workflow to analyse every single collision produced in the LHC

- **WP4: Education Programmes and Outreach**

- foster and train computing skills in the next generation of high energy physicists



Disclaimer:

The material presented in this talk is a general overview of the various recent efforts across the 15 tasks between ATLAS (WP2) and CMS (WP3). WP1 and WP4 are not covered due to time constraints, despite all the significant work carried out in those tasks. The contributions span substantially beyond what is presented here.

WPI & WP4: Infrastructure, Algorithms and Theory & Education Programmes and Outreach

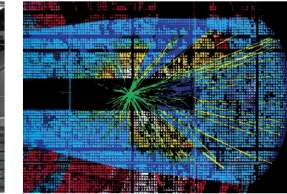
<https://nextgentriggers.web.cern.ch>
NGT 1st Technical Workshop

WPI = Infrastructure, Algorithms and Theory

- Task 1.1: Hardware and services for large scale NN optimisation and training, and physics simulation
- Task 1.2: Fast inference of complex network architectures on LHC online systems
- Task 1.3: Hardware-aware AI optimisation
- Task 1.4: Tensor Networks for Quantum Systems
- Task 1.5: New computing strategies for data modelling and interpretation
- Task 1.6: New Physics scenarios and Standard Model properties as trigger benchmarks
- Task 1.7: Common software developments for heterogeneous architectures



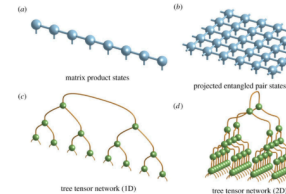
T1.1: Hardware and services for large scale NN optimisation and training, and physics simulation



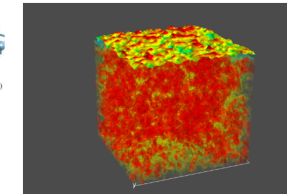
T1.2: Fast inference of complex network architectures on LHC online systems



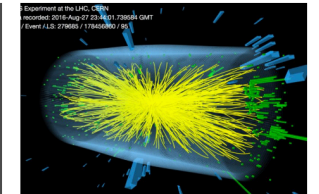
T1.3: Hardware-aware AI optimization



T1.4: Tensor Networks for Quantum Systems



T1.5: New computing strategies for data modeling and interpretation



T1.6: New Physics scenarios and Standard Model properties as trigger benchmarks

WP4 = Education Programmes and Outreach

- Task 4.1: Exchange Programmes and Outreach
- Task 4.2: The STEAM Programme (Software Training, Education, and Advanced Modules)

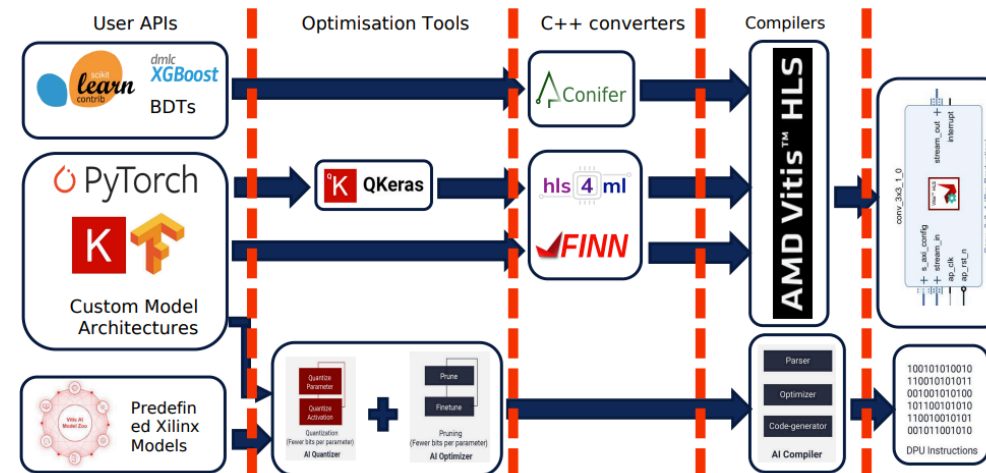
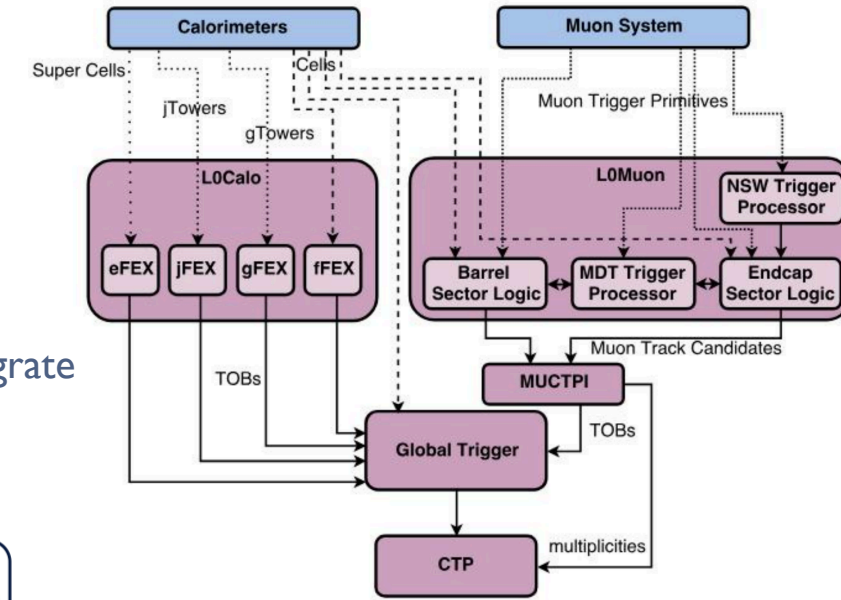
Strong synergies and support for WP2 (ATLAS) and WP3 (CMS) tasks!



ATLAS Hardware Level-0 (L0) Trigger

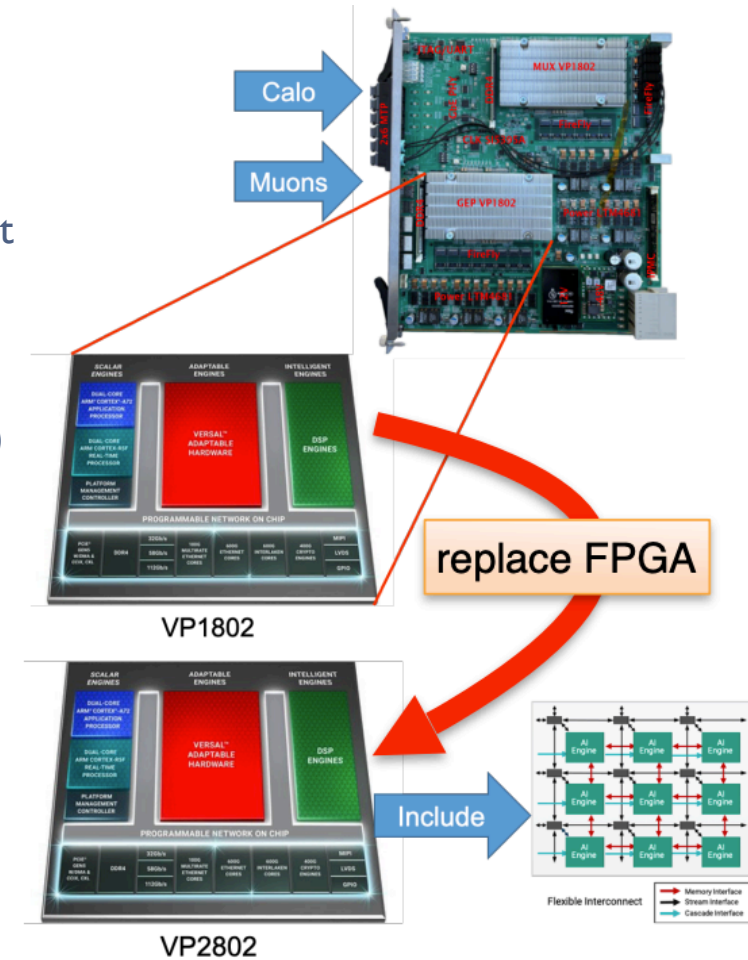
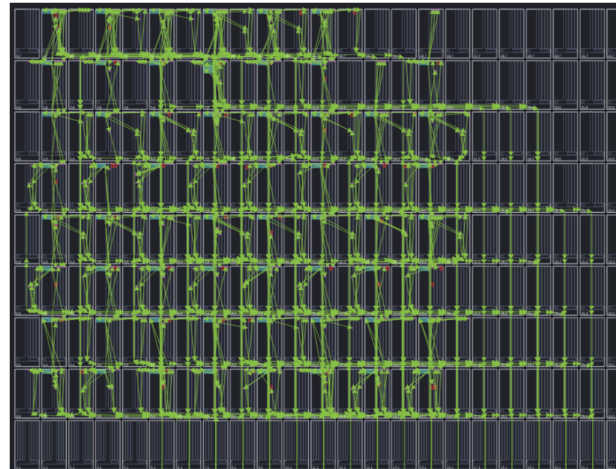
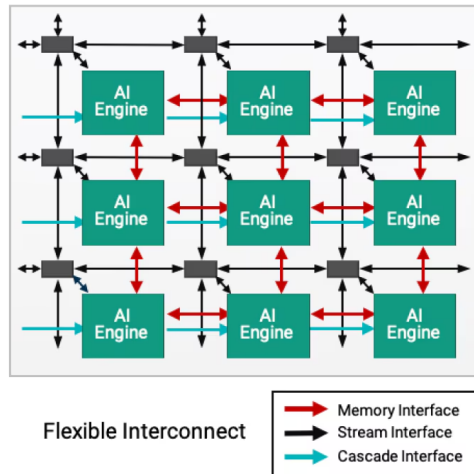
ATLAS (Task 2.1): Optimal Real-Time Event Selection in the Global Trigger System

- Explore novel **machine learning (ML)** reconstruction within the Level-0 (L0) Global Trigger:
 - critical system responsible for reconstructing in real-time all events
- Use-cases for novel ML algorithms for L0-Global Trigger:
 - Boosted-Decision-Trees (BDTs) for L0 e/γ trigger selection
 - convolutional neural network (CNN)-based pileup rejection and large-R jet tagging
- Investigation of different pipelines to bring ML approaches onto **FPGAs**
 - developing a common framework for ML optimisation for FPGA development to integrate within the Global Event Processor (GEP)



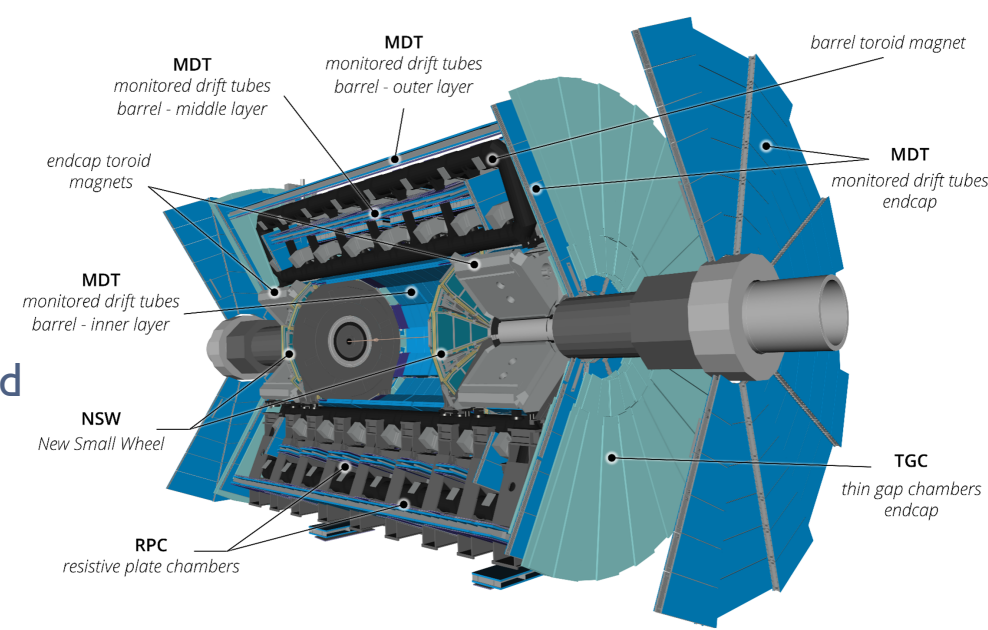
ATLAS (Task 2.1): Optimal Real-Time Event Selection in the Global Trigger System

- Evaluation of new **industry technologies**:
 - L0-Global Trigger deals with the ATLAS hardware trigger constraints by deploying a single hardware board called Global Common Module (GCM):
 - hosts a Versal Premium device (AMD VPI802) due to the big amount of input/output
 - however, provides limited resources for ML algorithms
 - Incorporation of **AI engines (AIE)** = dedicated vector processors
 - testing a prototype GCM with pin-compatible FPGA with AI engines (AMD VP2802)



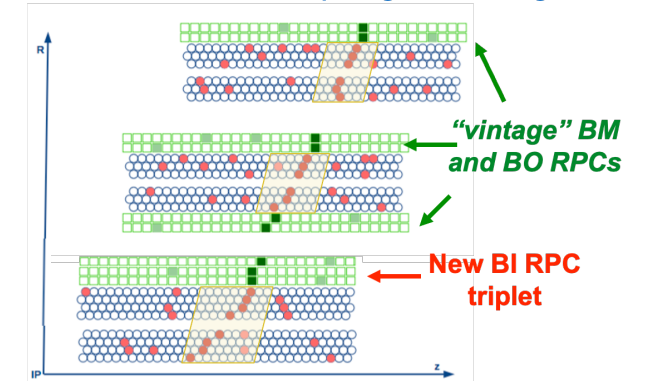
ATLAS (Task 2.2): Enhancing the Level-0 Muon Trigger

- Main objective is to design and implement **new algorithms** to extend the Level-0 (L0) Muon Trigger
 - focus on improving robustness and overall acceptance
- Baseline L0 Muon Trigger in the ATLAS barrel relies on resistive plate chambers (RPCs) for best selectivity and performance
 - improve trigger robustness in case of reduced RPC performance
- Take advantage of muon detector Phase-2 upgrades for HL-LHC:
 - precision measurements from monitored drift tube (MDT) chambers
 - improved efficiency with additional layer of RPCs in barrel
 - new electronics including modern FPGAs for trigger and readout



Hit Extraction

RPCs provide seeds to identify MDT hits from a muon & set up segment fitting

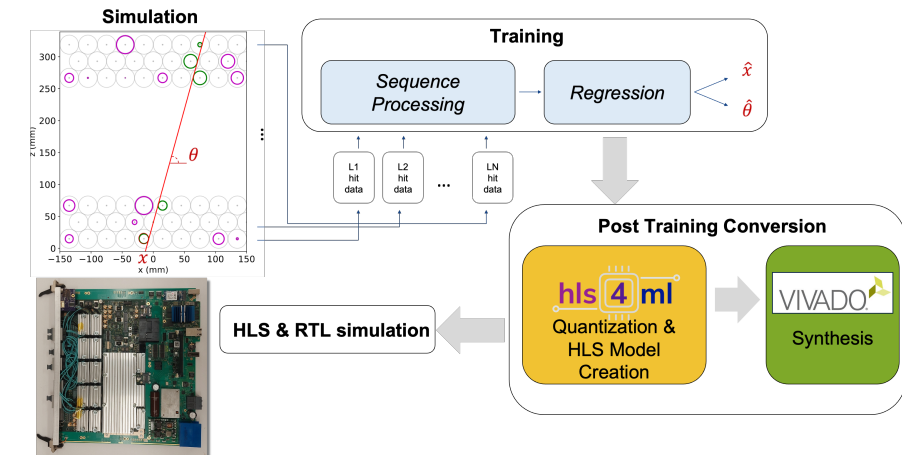
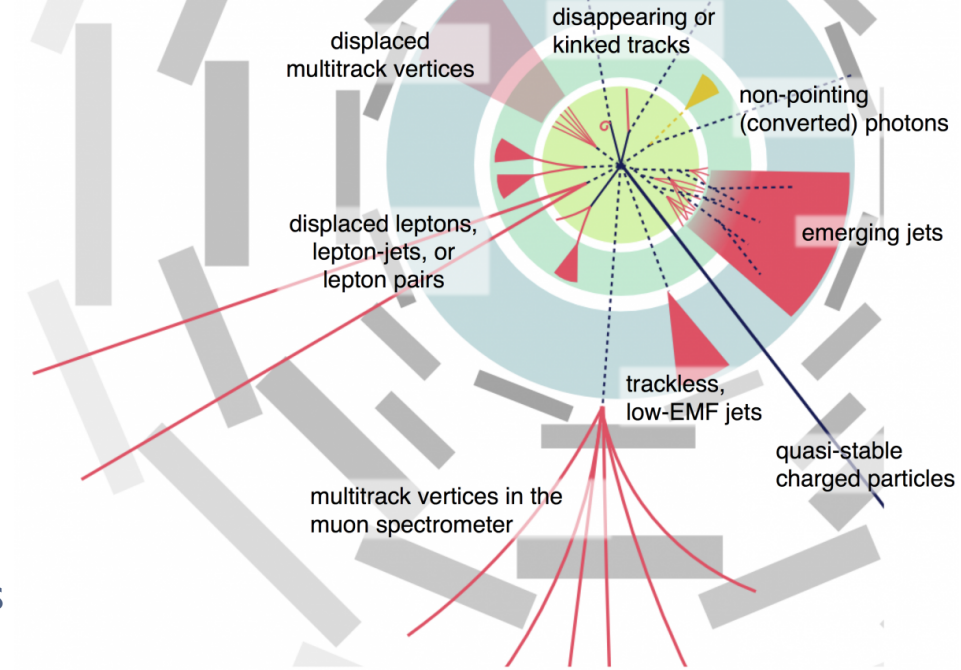


Pattern recognition algorithms to identify the regions of interest with only MDT hits

ATLAS (Task 2.2): Enhancing the Level-0 Muon Trigger

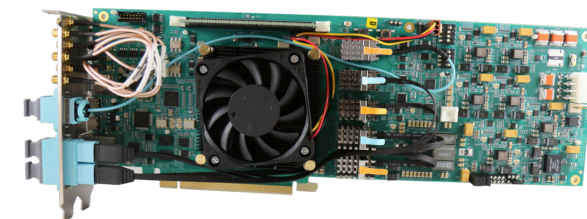
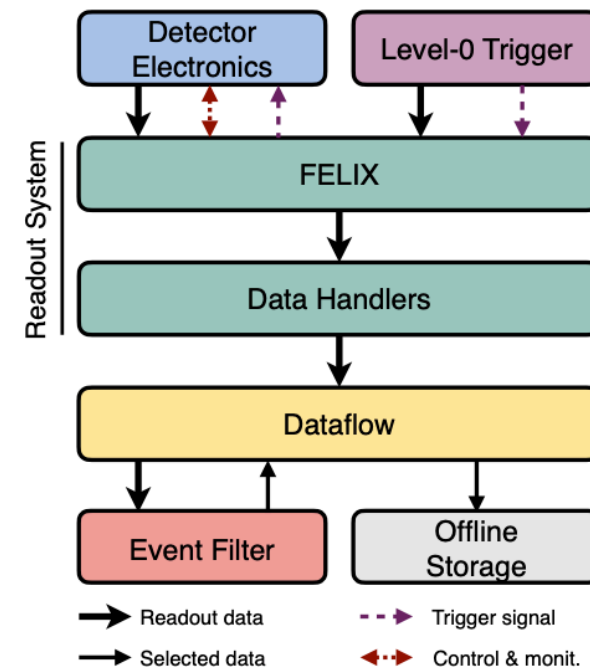
Algorithm development for the Level-0 (L0) muon trigger:

- Baseline design does not target exotic signatures from long-lived particles
 - developing new algorithms to target **displaced muons**
 - from decays of long-lived particles, predicted by a number models of new physics ([J. Phys. G 47 \(2020\) 090501](#))
 - other signatures to explore in the future (e.g. slow moving or highly ionising particles, high muon multiplicity, nearby muons, ...)
- Exploration of approaches with **machine learning (ML)**:
 - investigating pattern recognition algorithms to identify the regions of interest using MDT hits (incl. CNNs with detector granularity)
 - exploring algorithms more suitable for sparse data such as recurrent (RNNs) or graph neural networks (GNNs) for pattern recognition
 - ML methods for momentum estimation



ATLAS (Task 2.3): High Throughput Data-Collection

- Focus on improving and optimising the baseline **readout capabilities** of ATLAS:
 - interface between the detector and trigger electronics and a commercial computer network = 5 TB/s data throughput from 15k optical links
 - Front-End Link eXchange (FELIX) cards/hosts for packet routing (readout & control and monitoring), trigger & clock signal distribution
 - Data Handler (DH) for detector-specific packet processing and aggregation into bigger fragments
- **Network technologies** (TCP/IP vs RDMA; converged RoCEv2 vs Ultra-Ethernet)
 - progress on the development of a novel network library developed within FELIX to support these
- Performed FELIX and DH performance measurements in representative scenarios to determine CPU requirements
- Evaluation of **alternative readout architectures**, including exploring market solutions in software and nonstandard hardware:
 - FELIX & DH on the same host (server with 2x96-core CPUs)

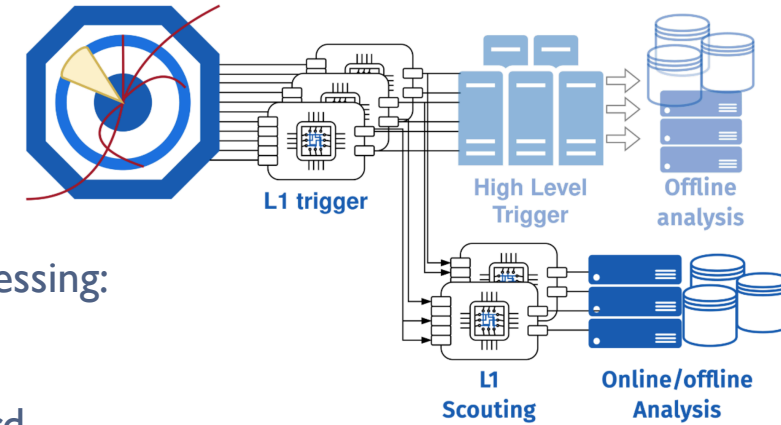


FELIX (FLX-182): Custom-designed AMD Versal Prime (VM1802) FPGA, PCIe Gen5



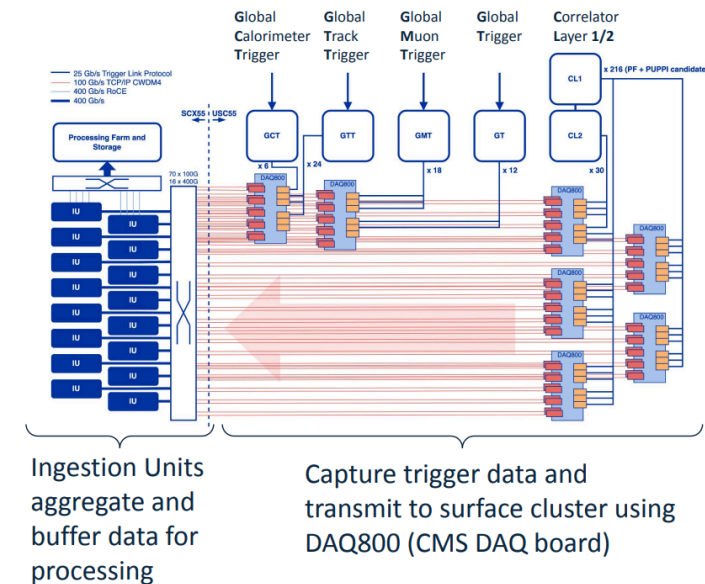
CMS Hardware Level-1 Trigger (L1T)

CMS (Task 3.5): L1 Scouting for HL-LHC



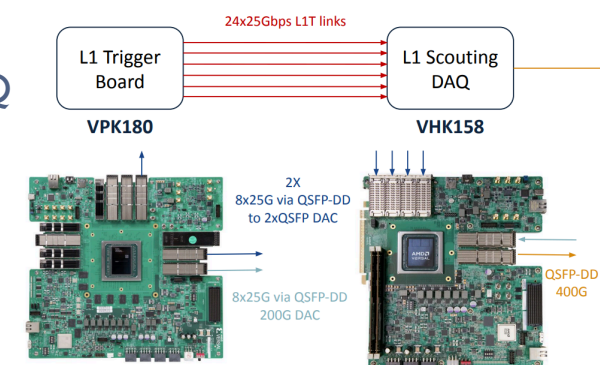
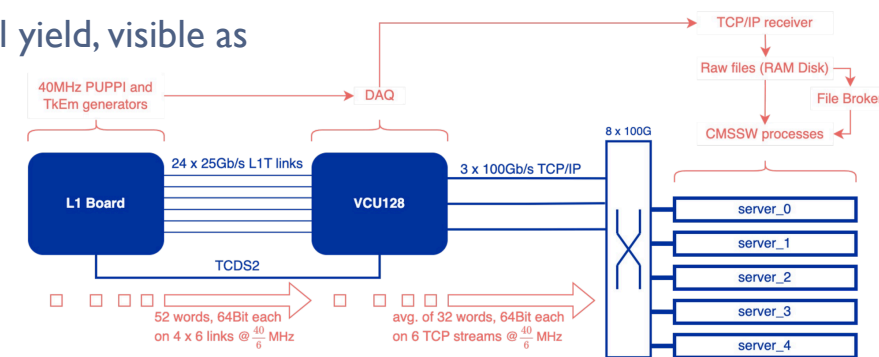
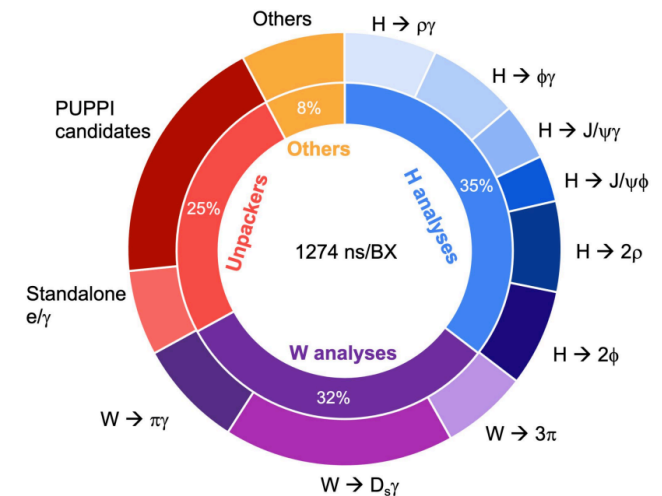
- **Scouting** = reduced data format overcoming the limits of the conventional data processing:
 - first implemented @ HLT → see recent HLT-scouting paper ([CMS-EXO-23-007](#))
 - extends the phase space that is unexplored, unreachable or inefficient with standard trigger strategies by accessing lower trigger thresholds
- **LIT data scouting** (LIDS): capture LIT objects reconstructed at 40 MHz and run physics analysis on-the-fly to fully exploit the potential in Phase-2
 - use of dedicated FPGA-based boards collecting LIT objects via optical links
- Difference wrt. HLT-scouting = runs at 40 MHz (but no full detector readout)
 - plan is to leverage improved object reconstruction quality at LIT in Phase-2
 - e.g. tracking, particle flow (PF), pileup per particle identification (PUPPI), ...
- **LIDS Run-3 demonstrator** has been employed in the LIT since the start of Run-3 as proof-of-concept → collecting and analysing data at 40 MHz in 2024 ([CMS-DP-2024-056](#))

Baseline LIDS architecture for Phase-2



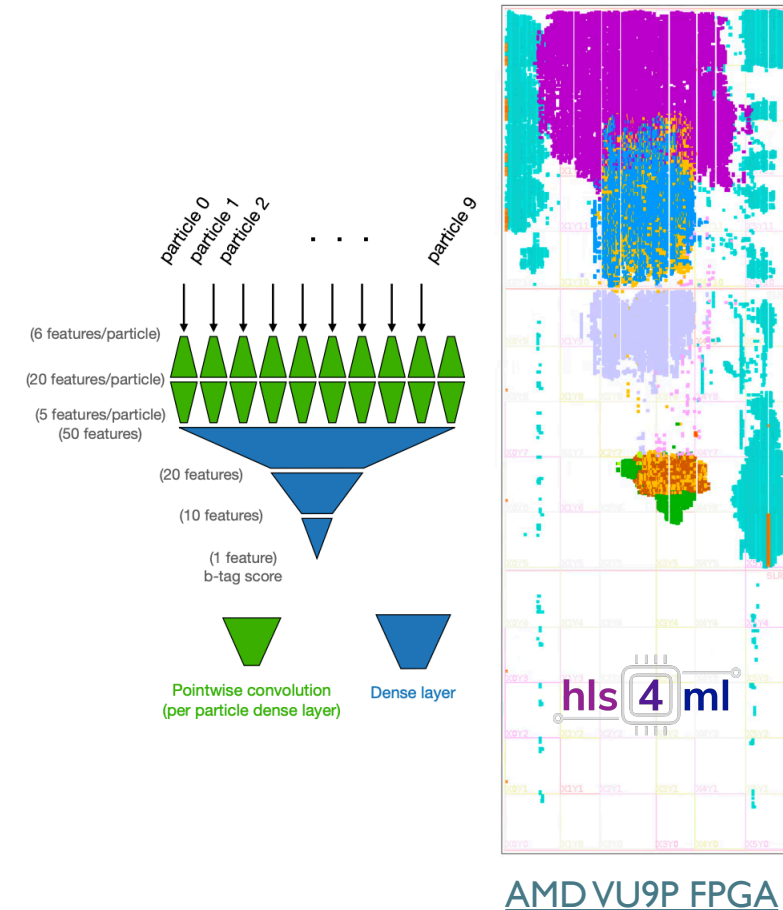
CMS (Task 3.5): L1 Scouting for HL-LHC

- NGT goal is to accelerate baseline **physics studies** and explore potential for going beyond with more **advanced algorithms** and more L1T input
- Built a **prototype** for the baseline LIDS system for Phase-2 in order to develop DAQ and real-time processing components ([CMS-DP-2024-096](#))
 - Several **prototype analyses** implemented, featuring soft final states with a small signal yield, visible as narrow peak over smooth background (uncaptured by standard L1T)
- Test **alternative architectures** for processing on baseline hardware:
 - detector-to-surface network protocols (e.g. Ultra-Ethernet)
 - more performant hardware e.g. next generation FPGAs (56 Gbps+) with RoCE
 - implementation of large low-latency buffers for online processing
 - e.g. SSDs or compute express link (CXL)-enabled “memory lake” accelerators (Micron)
- Developing novel approaches for processing LIDS data with **beyond-baseline hardware** for DAQ
 - explore accelerators e.g. GPUs, FPGAs and AMD AI engines (AIE) for offloading of complex reconstruction/analysis tasks and ML models (incl. Alpaka)
 - custom DAQ boards based on newer technologies (e.g. AMD Versal VHK158)
 - develop effective data-compression algorithms for L1T scouting data



CMS (Task 3.6): Practical Real-Time AI for LI Trigger and LI Scouting

- Processing at LIT uses **dedicated algorithms** for physics reconstruction (incl. ML)
 - integration of ML in the LIT, anticipating 25 billion inferences per second in Phase-2
 - sub-microsecond latency ML on FPGAs using high-level-synthesis (HLS) for machine learning ([hls4ml](#)) and conifer tools for NNs and BDTs
- **Optimal reconstruction** of high-level (particles, energy sums, jets, tracks) from low-level (detector hits) information to trigger on hard-to-select signatures
 - plan is to leverage improved object reconstruction quality at LIT in Phase-2
 - e.g. tracking, particle flow (PF), pileup per particle identification (PUPPI), ...
 - use NN architectures to propagate dense information downstream compactly where it can be used to select events
- **ML-Operations (ML-OPS)** needed to run a hardware trigger with many ML-models
 - ML lifecycle: all steps from dataset creation, model training, model repository, synthesis, emulation
 - MLFlow and Gitlab CI applied to jet tagging and cluster ID (see next slide)
 - facilitating widespread use of NNs in trigger → robust and sustainable

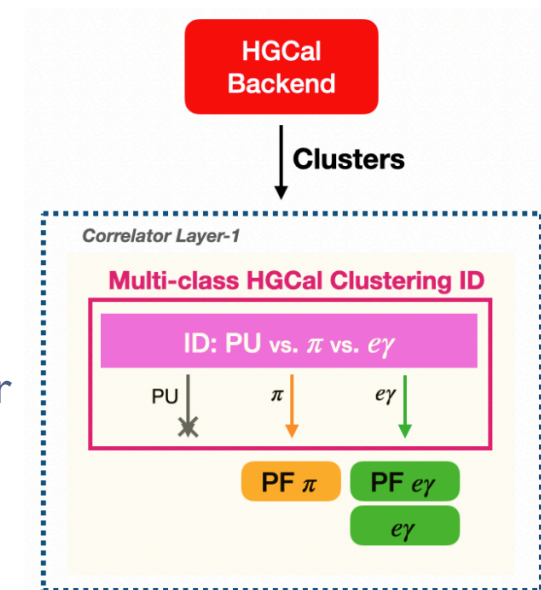


CMS (Task 3.6): Practical Real-Time AI for LI Trigger and LI Scouting

- Dedicated applications of ML in LIT in Phase-2, with the NGT goal to extend these practices further to other high-level-objects as well as to topological signatures targeting specific final states:
 - **jet tagging**: enhancing baseline jet tagging at LIT with NN ([CMS-DP-2022-021](#)) with correlator jet tagging based on the DeepSets architecture
 - classifying 8 classes: light jet (uds), gluon jet, b jet, c jet, $\pm\tau_h$, electron, muon
 - to improve acceptance for final states like $HH \rightarrow b\bar{b}\tau\tau$, exploiting combined b-tagging and tau-tagging information

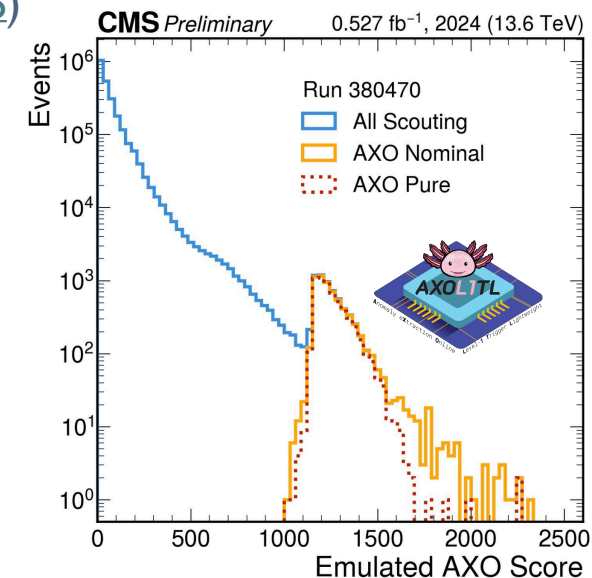
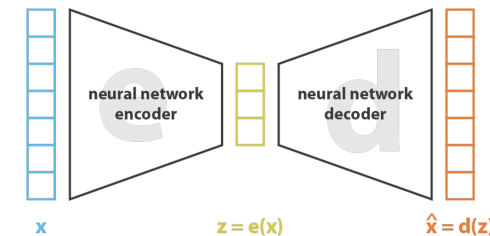


- **cluster ID**: lightweight fast identification of pileup clusters with BDTs in the HGCal for background rejection ([CMS-DP-2024-098](#))
 - improve categorisation help to enhance efficiency for soft e/ γ clusters
→ allowing triggering (scouting) on low transverse momentum electrons

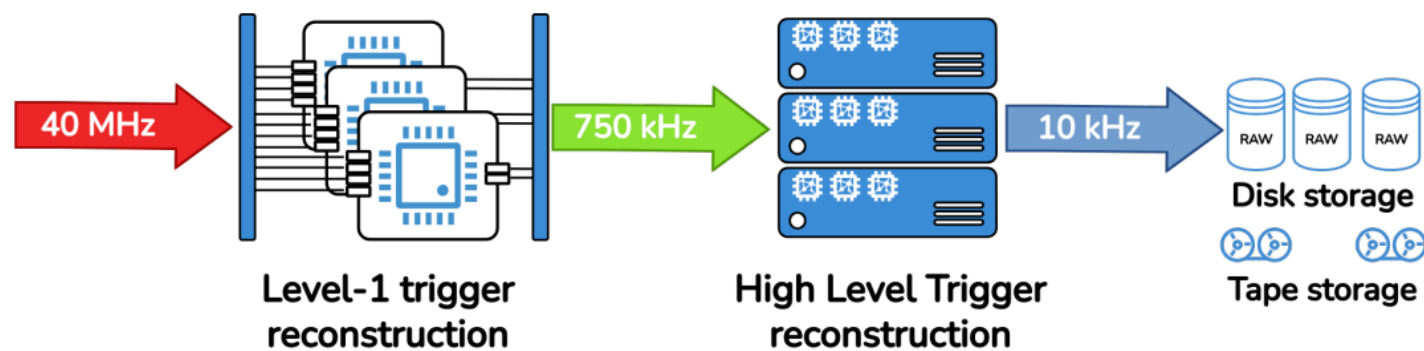


CMS (Task 3.7): LI Scouting Data Compression for Efficient Data Acquisition and Anomaly Detection

- **Anomaly detection @ LHC** = generalising NP searches to a large variety of BSM models at once
 - identifying rare events in datasets which deviate significantly from the majority of the data and do not conform to “normal” behaviour
 - such behaviour is learnt through neural-networks (NNs): AutoEncoders (AEs) → Variational AEs (VAEs)
- **Ultra-fast anomaly detection @ LIT** already deployed during LHC Run-3 (100 fb⁻¹, 18% pure rate):
 - AXOLITL: Anomaly eXtraction Online Level-I Trigger aLgorithm ([CMS-DP-2023-079](#) and [CMS-DP-2024-059](#))
 - CICADA: Calorimeter Image Convolutional Anomaly Detection Algorithm ([CMS-DP-2023-086](#))
- NGT focus to **improve AXOLITL** during Run-3 and for Phase-2:
 - advancing multiple aspects of the project in Run-3: physics analysis, model development and operational automation
 - e.g. design more robust model based on contrastive learning techniques (SSL)
 - develop an upgraded model tailored to the Phase-2 LIT system
 - e.g. incorporate new inputs and improved reconstructed objects
- Design novel **point-cloud-based** anomaly detection algorithms:
 - use all particles reconstructed from LIT as input

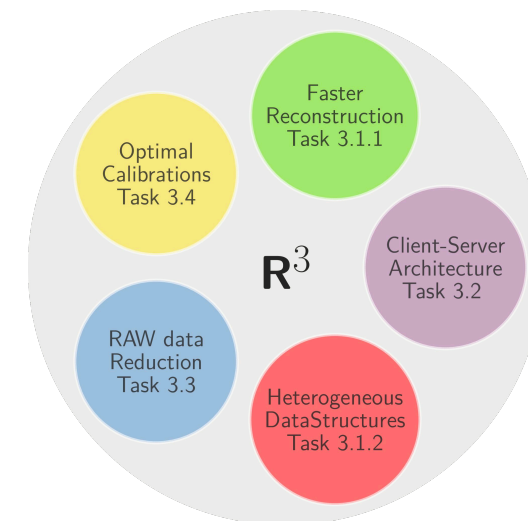
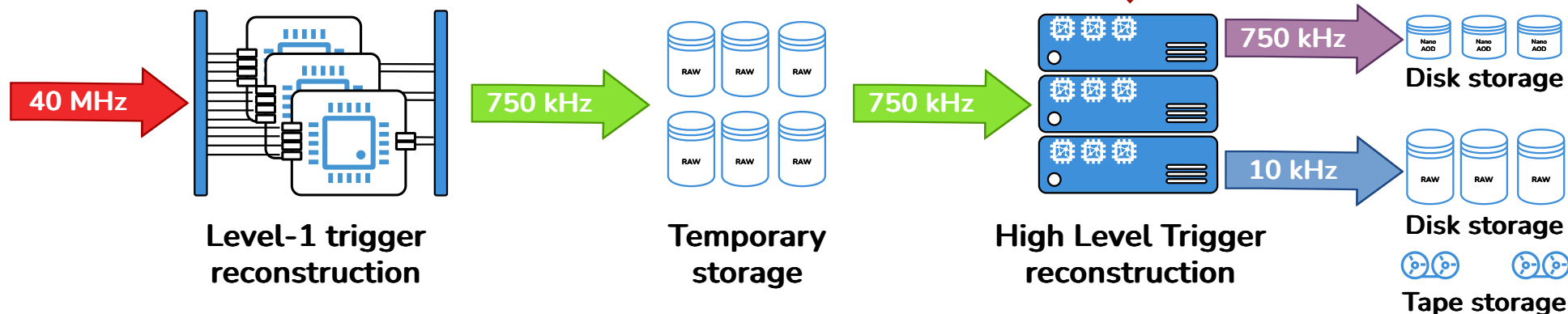


CMS Software High-Level Trigger (HLT)



CMS (NGT-HLT): Real-time Reconstruction Revolution (R^3)

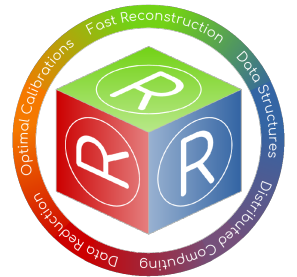
- Goal of R^3 project = **overcome the main limitations of the HLT**
 - quality of the online reconstruction → limited by the processing capacity (timing/throughput)
 - HLT output rate → limited by the storage capacity and processing power of the offline computing infrastructure
- Ambitious plan is to have offline-like quality reconstruction (incl. calibrations) at the HLT...
- ...by developing **modern and GPU-friendly algorithms & data structures for faster online reconstruction, optimal calibrations, data size reduction**



NextGen

CMS (Task 3.1.1): R³ Faster Reconstruction

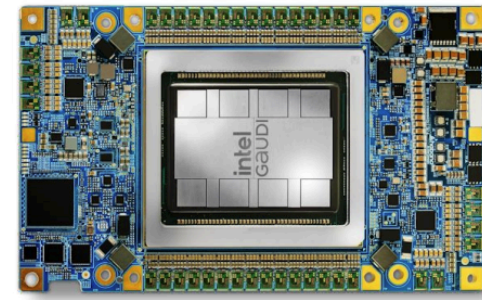
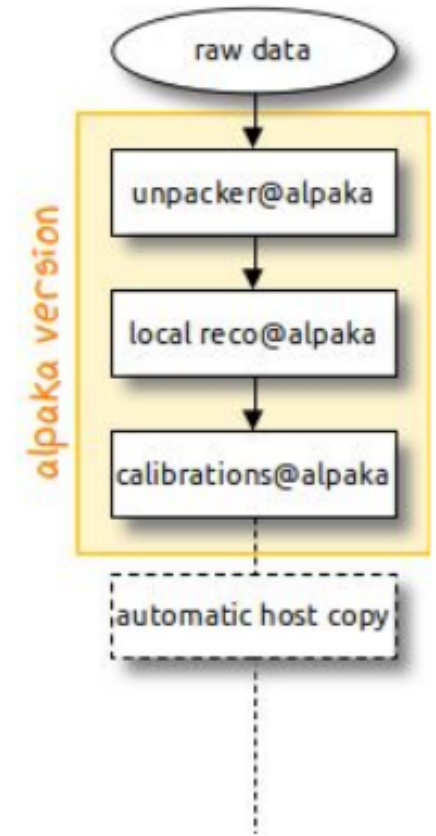
- Successful experience of **heterogeneous reconstruction** in Run 3 has shown that it is possible to improve the physics quality and throughput of selected physics objects (e.g. Patatrack pixel tracks) on GPUs
- Ambitious R³ project aims to:
 - modernise the system by using novel methodologies and leveraging **heterogeneous** compute resources and **ML-driven** techniques
 - redesign the most important physics objects (muons, electrons, photons, taus, jets, missing transverse momentum and particle flow-PF event description)
 - perform **offline-like quality** event reconstruction at the **full LIT input rate** (500 - 750 kHz)
- Measurements and extrapolations for the HLT reconstruction for baseline Phase-2:
 - Assume 500 kHz input rate for Run-4 (2030), 750 kHz for Run-5 (2035)
 - Assume at least 50% code runs on GPU by Run-4 and 80% on GPU by Run-5
 - Assume flat +20% improvements in performance per year from hardware and cost (CPUs & GPUs)
- Requires a speedup of the baseline online reconstruction (ported from Run-3) by **2x** by Run-4
 - To achieve ambitious R³ goal → requires speedup by an additional factor of **2x - 8x** (!)



NexTGen

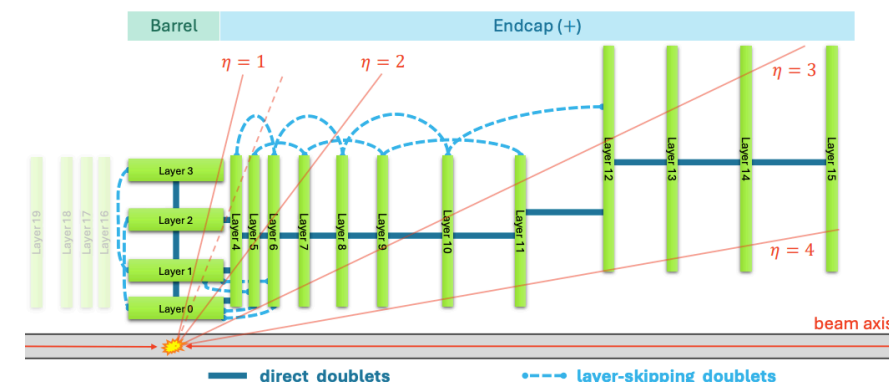
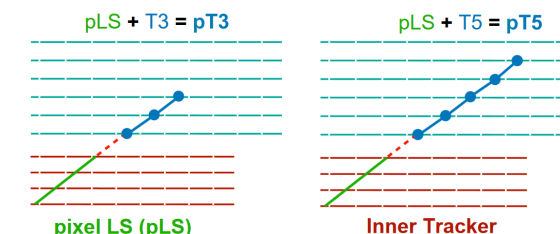
CMS (Task 3.1.1): R³ Faster Reconstruction

- Expanding further on Run-3 **GPU improvements** (also baseline of the Phase-2 programme):
 - Currently offloading to GPU ~35% of HLT reconstruction with a ~50% speedup compared to CPU only
 - Heterogeneous Patatrack (Alpaka) pixel tracks and ECAL, HCAL and Particle Flow (PF)
 - Migration from CUDA to Alpaka = performance portability library allows a single source to be built for and run on: CPUs, GPU (experimental support for FPGAs)
- Pushing further on heterogeneous computing and integration of **accelerators** → investigation of alternative accelerator platforms:
 - FPGAs for serial reconstruction
 - dedicated hardware for ML workflows → AI engines (AIEs)
 - accessible over a high-speed RDMA network



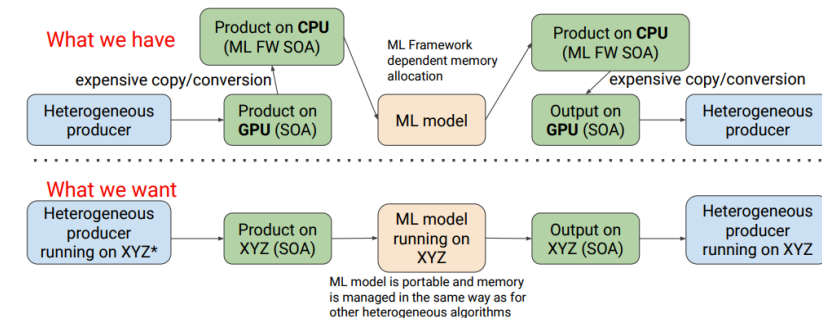
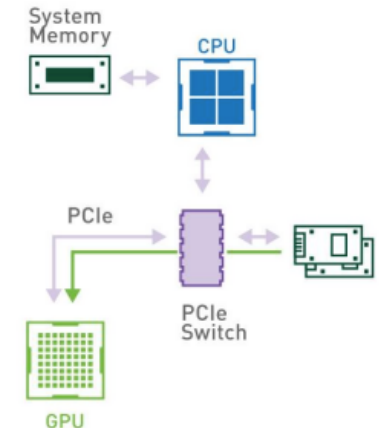
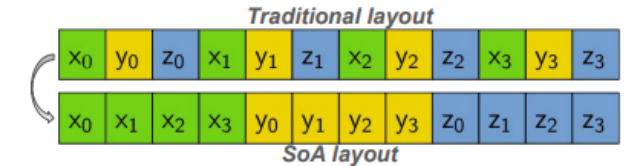
CMS (Task 3.1.1): R³ Faster Reconstruction

- Development to reach this goal is focusing on more efficient **traditional** and **ML-based reconstruction algorithms** and **data-structures** to speed up tracking, clustering and pattern recognition
 - combinatorial nature of pattern recognition in tracking → superlinear increase of computational complexity with input
- Explore **innovative techniques** from the many on-going baseline Phase-2 developments such as:
 - Line Segment Tracking (LST)** as an alternative to Kalman filters to improve reconstruction performance and flexibility
 - Exploit the new tracker to improve timing and extend physics acceptance (especially for displaced tracks)
 - Designed for parallelisation → able to leverage heterogenous computing on GPUs (Alpaka)
 - Machine learning parts to improve pattern recognition
 - AI/ML-driven solutions** to support novel methods for complex reconstruction challenges:
 - The Iterative CLustering (TICL) reconstruction (includes DNN super-clustering in HGCal)
 - Expand fast GPU inference to reduce latency and improve efficiency
- On-going **optimisation** work on:
 - HLT (Alpaka) pixel tracks
 - MC truth information with SimDoublets/SimNtuplets
 - optimiser of doublet cut parameters
 - HLT standalone (L2) and tracker (L3) muon reconstruction



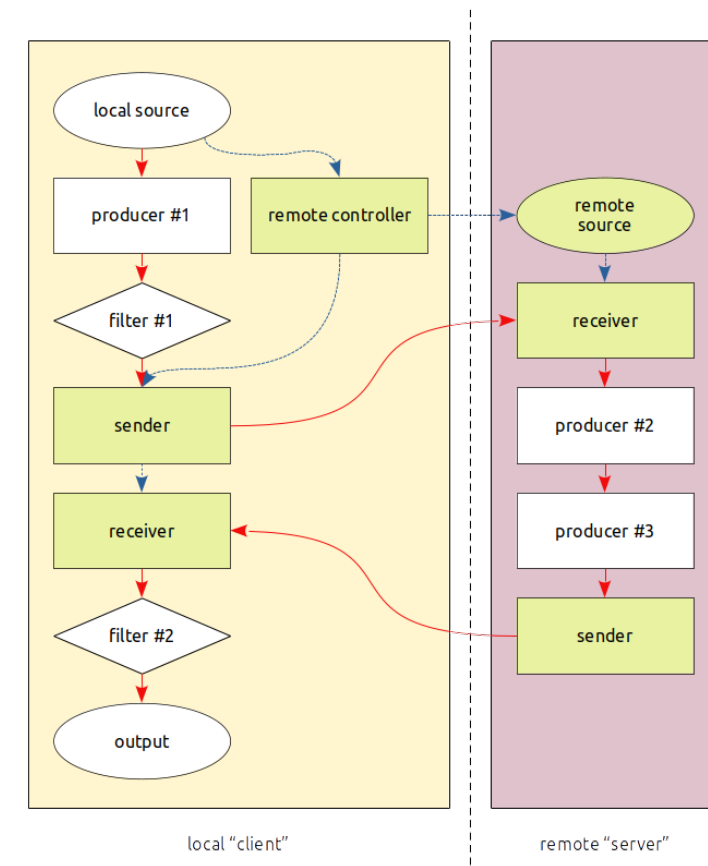
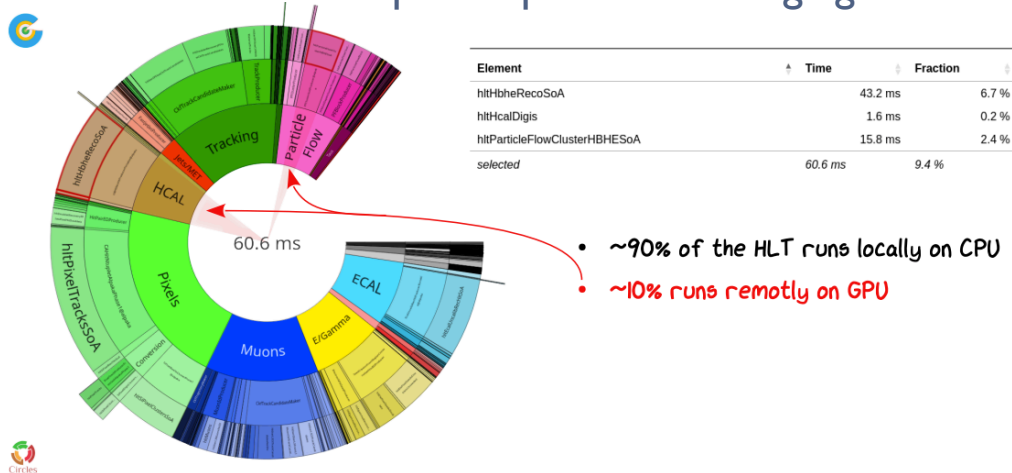
CMS (Task 3.1.2): R^3 Optimised Data Structures for Heterogeneous Platforms

- Algorithms should be designed in tandem with the corresponding data structures:
 - data structures in heterogeneous environments might become the bottleneck when the cost of sequential copies and conversions are not negligible
 - efficient data structures** are essential for optimising performance in heterogeneous computing environments
- Efficient memory access patterns → coalesced memory access for GPUs
 - Struct of Arrays (SoA)** layout is beneficial when one needs to perform operations on some fields for all elements concurrently
- NGT work on extending the existing SoA approach already used in Run-3
- Flexibility, maintainability and efficiency through **standardisation**:
 - generic and flexible SoA compositions for efficient memory usage
 - ensure compatibility with C++20 standard
- Seamless **integration** with other developments:
 - direct interface to heterogeneous ML models
 - remote offload through message passing (Task 3.2)



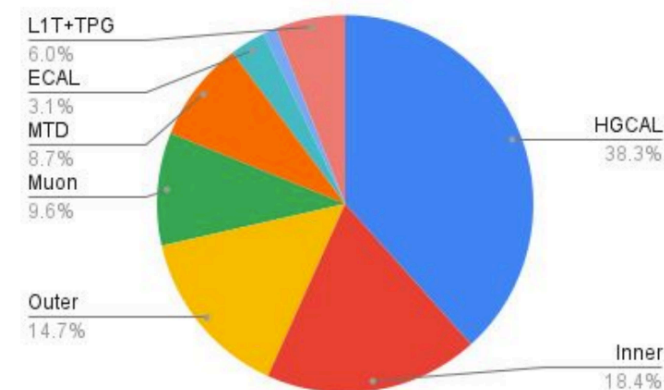
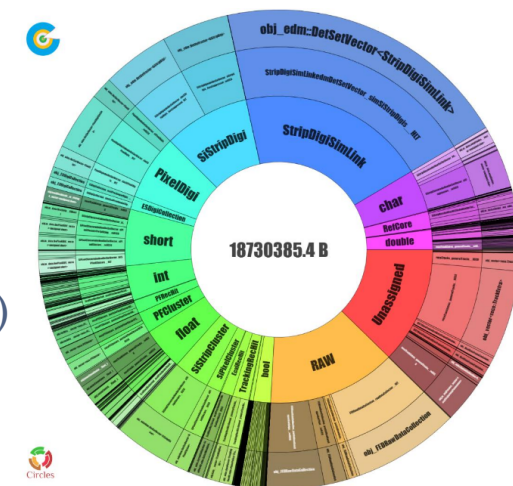
CMS (Task 3.2): Evolving the CMS Experiment Software into a Client-service Distributed Application for HLT

- CMS software (CMSSW) is modular, parallel, heterogeneous application framework for CMS data acquisition, reconstruction, processing and analysis, built around an Event Data Model (EDM) based on C++ and Python
- Plan to design and implement CMSSW as a **distributed application** (with minimal impact on the existing code base)
 - in order to exploit **remote accelerators**
 - e.g. offload GPU reconstruction to remote accelerators
- **Prototype** based on the Message Passing Interface (MPI) offloading part of the HLT HCAL reconstruction to a separate process leveraging GPUs:



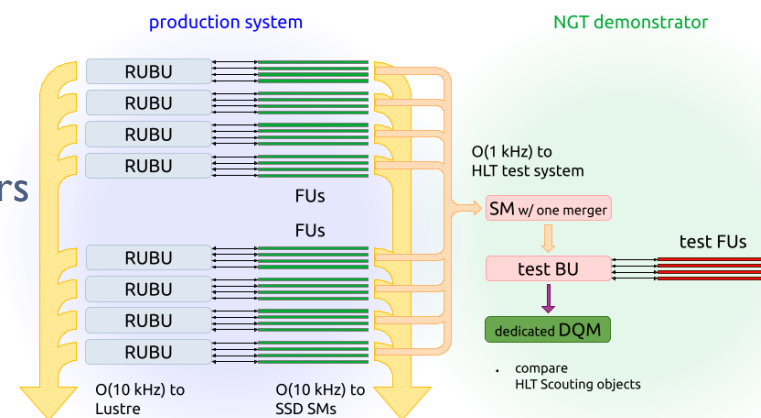
CMS (Task 3.3): Reduction of the RAW Data Size for HLT

- R³ project aims to run an offline-like reconstruction on **all events** accepted by the LIT (500 - 750 kHz)
 - Phase-2 event size of CMS is expected to be around 10 MB/event (new detectors and higher pileup)
 - not possible to store all events accepted from the LIT in RAW data format (7.5 TB/s)
 - goal is to increase the maximum available bandwidth by reducing the size of events saved by CMS
- **Lossy compression**, replacing raw data with **high-level physics objects** (e.g. muons, electrons, jets, tracks)
 - evolution of our current HLT-scouting data format
 - very strong compression (around factor $\times 100$) necessary to store all 750 kHz
 - limited possibility to re-reconstruct objects with newer algorithm or calibrations
- **Lossy compression**, replacing raw data with **low-level physics objects** (by approximation)
 - example: replacing raw data with reconstructed hit positions and energies
 - extension of RAW' compressed data-format currently used in heavy-ion collisions in CMS
 - limited compression (below $\times 10$) e.g. reduction $\sim 30\%$ in Run 3 proton-proton data
 - will allow to re-reconstruct high-level objects with newer algorithm or calibrations
- **Lossless compression**:
 - testing and benchmarking new compression algorithms (LZMA, ZLIB, ZSTD, LZ4)
 - using physics objects (eg. tracks) to improve the raw data compression



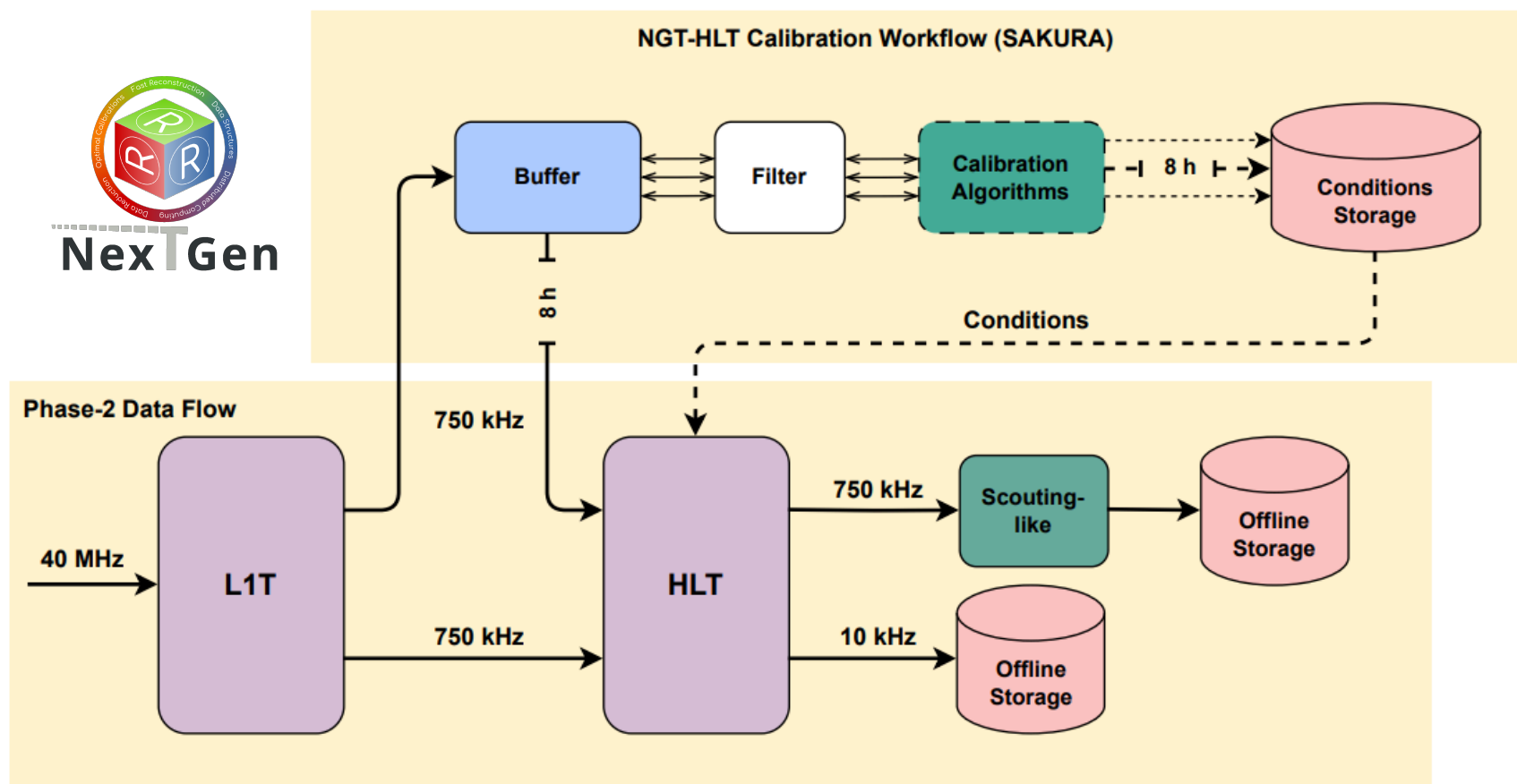
CMS (Task 3.4): Optimal Calibrations for HLT

- SAKURA (Speedy Alignment & Calibration Upgrade for the R³ Algorithms) project aims to design **accelerated calibration workflows** at HLT to achieve the same accuracy as the prompt calibrations for offline reconstruction
 - inspired by existing LS-based beamspot workflow at HLT and [LHCb real-time alignment and calibrations](#)
- Run optimised (fast) online calibrations and inject them for final HLT reconstruction:
 - **buffer all input data** from L1T (500 - 750 kHz) and apply updated calibrations
 - à la Run-3 Prompt Calibration Loop (PCL), buffering data over several (8 - 12) hours
 - storage write bandwidth: $500 - 750 \text{ kHz} \times 8.4 \text{ MB/evt} = 4.2 - 6.3 \text{ TB/s}$
 - estimated buffer size $\approx 200 - 300 \text{ PB}$ (incl. 50% safety factor)
 - versus today's [LHCb BigBuffer \(40 PB\)](#) & [ALICE EOS buffer \(150 PB\)](#)
- Run 3 (2025) target: deploy online a small-scale **demonstrator/prototype** deriving improved calibrations (see backup)
 - hardware chain in the process of being implemented as part of the DAQ system
 - including 2x30 TB SSDs (Micron 9400 PRO) mounted on DAQ storage manager (SM) node
 - for buffering $\sim 1 \text{ kHz}$ RAW data for 8 hours (bandwidth: $1 \text{ kHz} \times 1.3 \text{ MB/evt} = 1.3 \text{ GB/s}$)



CMS (Task 3.4): Optimal Calibrations for HLT

NGT R³ workflow including the conceptual design of SAKURA optimal calibrations:



Calibration Candidates

Run-3 Demonstrator	Phase-2
	ECAL Laser Transparencies
	Beamspace
	Tracker Bad Components
	HCAL Pedestals, Gains
Tracker Alignment (LG)	Tracker Alignment
Strips Particle Gains (G2)	
	ECAL Pedestals (G12), Pulse Shapes, Timing
⋮	⋮

NGT @ HLT (R³) Workflow

- top branch = on top of baseline Phase-2 HLT programme (bottom branch)
- output all L1T-accept data in a scouting-like format
- assume single chance to reconstruct these data with highest possible quality
→ requires optimal calibrations!

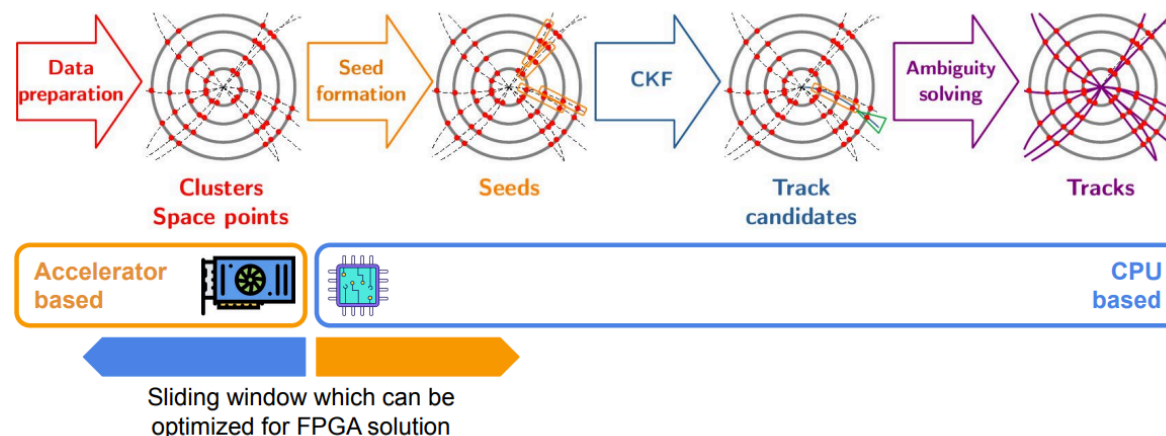
NGT developments also to be exploited at the baseline HLT branch!



ATLAS Software Event Filter (EF) Trigger

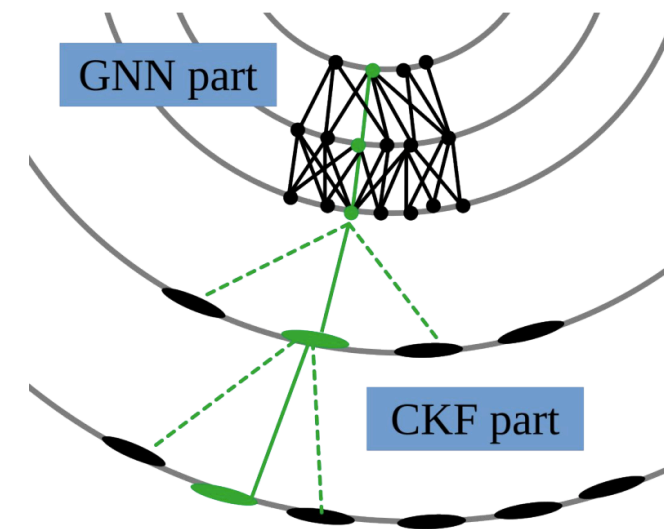
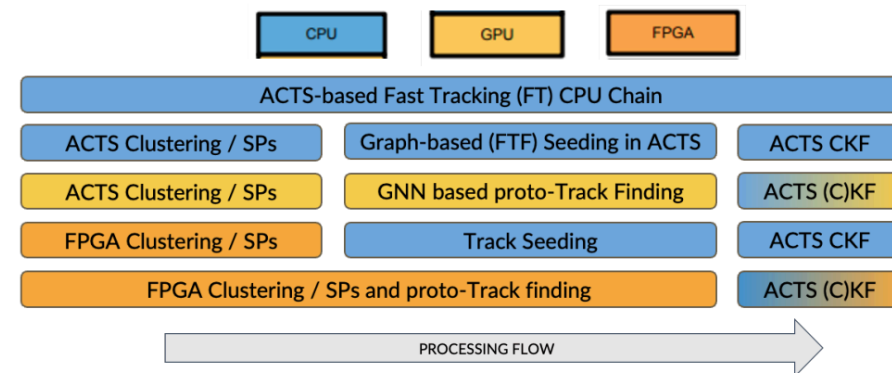
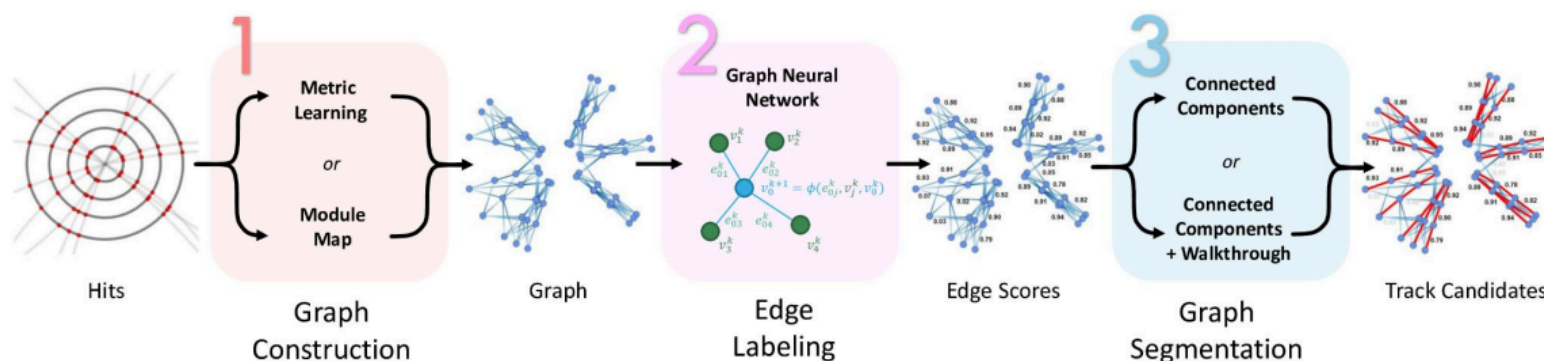
ATLAS (Task 2.4): Event Filter Tracking (ITk Reconstruction)

- Event Filter (EF) tracking = tracking algorithms running on flexible, heterogeneous commercial system
 - largest ongoing effort towards decisions on technology solutions for EF Tracking
- Development of an **algorithmic solution** for the EF track reconstruction for the Inner Tracker (ITk)
- Optimisation of physics and processing performance of the track reconstruction at the EF
- Looking for boundary between **CPUs**, **GPUs** and **FPGAs** for best computational and physics performance:
 - FPGA: major contribution on ITk data preparation on FPGA
 - CPU: development of ITk fast track reconstruction prototype for EF Tracking ([ATLAS-TDR-029-ADD-1](#))



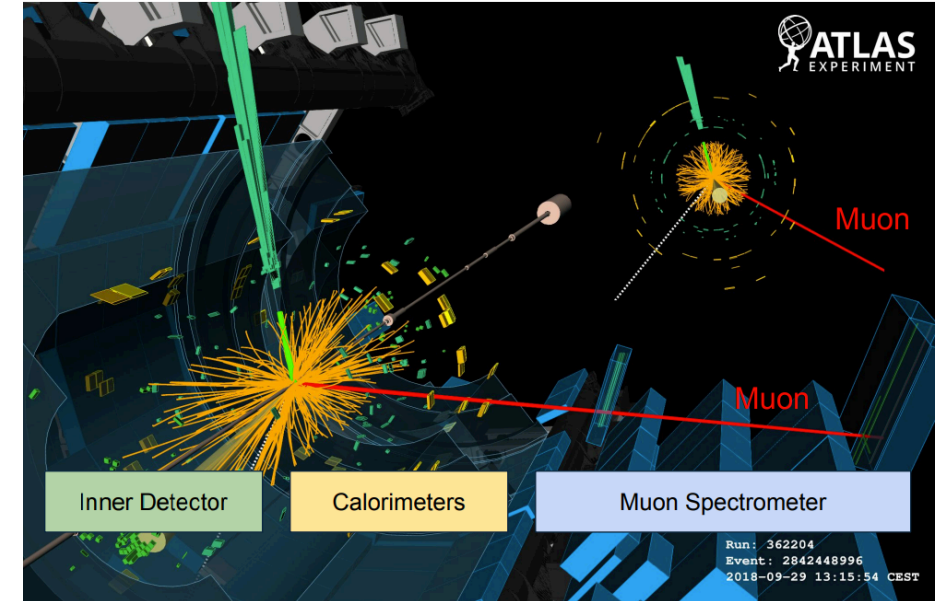
ATLAS (Task 2.4): Event Filter Tracking (ITk Reconstruction)

- Integration of [ACTS](#) (A Common Tracking Software), an experiment-independent track reconstruction **toolkit** in the ATLAS software for HL-LHC operations (Task 2.6)
 - CPU ACTS-based fast tracking chain fully integrated and on-going work on improving physics/CPU performance ([ATL-PHYS-PUB-2024-017](#))
- Exploring optimal classical and **ML techniques**, such as graph neural networks (GNNs) on GPUs as an alternative track finding approach ([ATL-PHYS-PUB-2024-018](#))
 - investigation of high performance inference frameworks such as Nvidia TensorRT for GNN inference
 - CUDA-based graph construction library of the offline GNN working group to facilitate integration into tracking frameworks



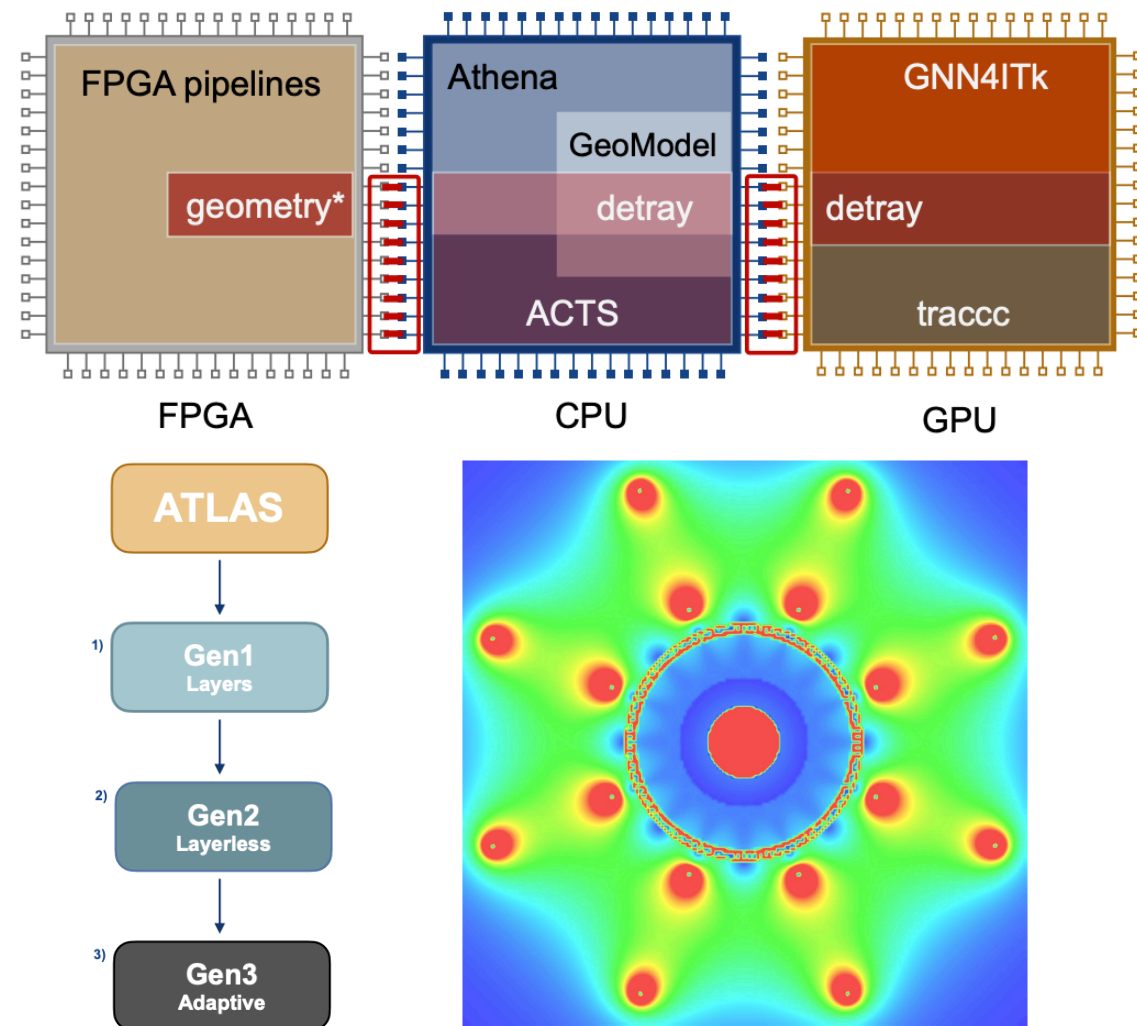
ATLAS (Task 2.5): Optimised Event Filter Muon Trigger

- Goal to improve the physics performance of the Event Filter (EF) **standalone muon track reconstruction** fully exploiting:
 - extended coverage of the L0 muon trigger (Task 2.2)
 - migration to novel ACTS tracking infrastructure (Task 2.6)
- Developing a novel segment finder based on a Hough Transform (HT) and a novel linearised χ^2 muon segment fit
- Developing novel **ML-based reconstruction** techniques (e.g. GNNs for sparse data) to improve on existing classical algorithm chain
 - reducing the likelihood of losing valid tracking seeds
 - improving overall tracking efficiency



ATLAS (Task 2.6): Common Tracking Event Filter Infrastructure

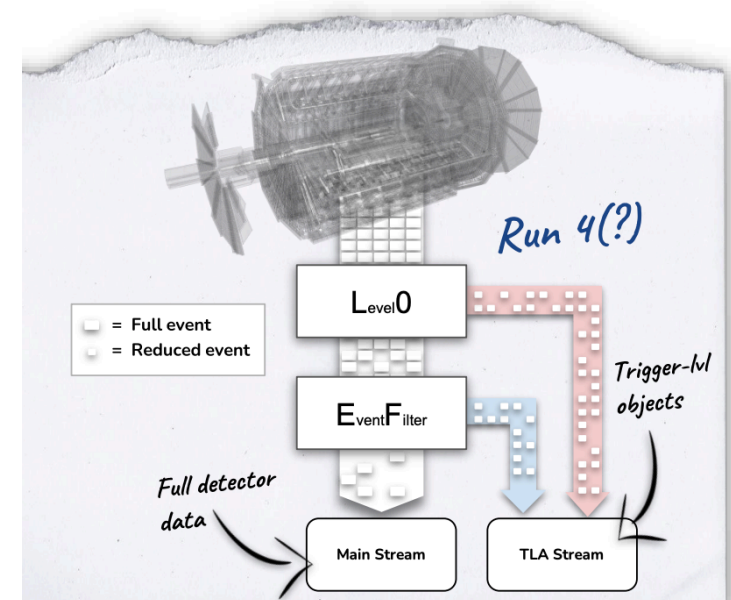
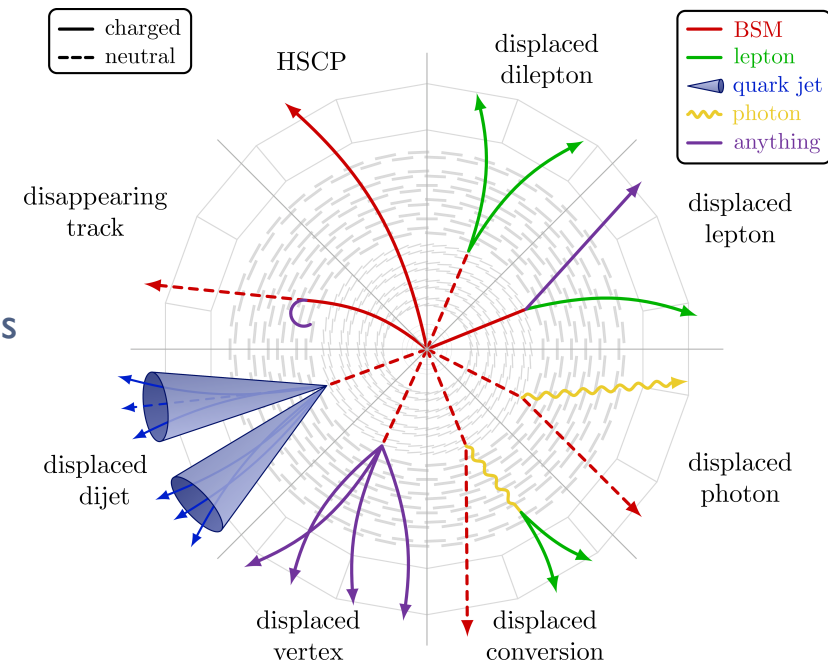
- Integration of **ACTS (A Common Tracking Software)**:
 - an experiment-independent track reconstruction toolkit in the ATLAS software for HL-LHC operations
- Provide **infrastructure support** for the various ATLAS Event Filter (EF) prototypes within ACTS:
 - including geometry, material, magnetic field and event data model support/integration (incl. ITk)
 - traccc (GPU reconstruction) integration into ACTS (CPU based toolkit)
- Improve and optimise integration of **GPU/CPU** (and eventual **FPGA**-based) demonstrators:
 - optimise data structures for heterogeneous pipelines
- Help with interfacing **ML pipelines** with ACTS



ATLAS (Task 2.7): Enhanced Reconstruction for Higher Level Event Filtering

Extension and optimisation of the Event Filter (EF) **trigger menu** with focus on physics application to increase the physics reach

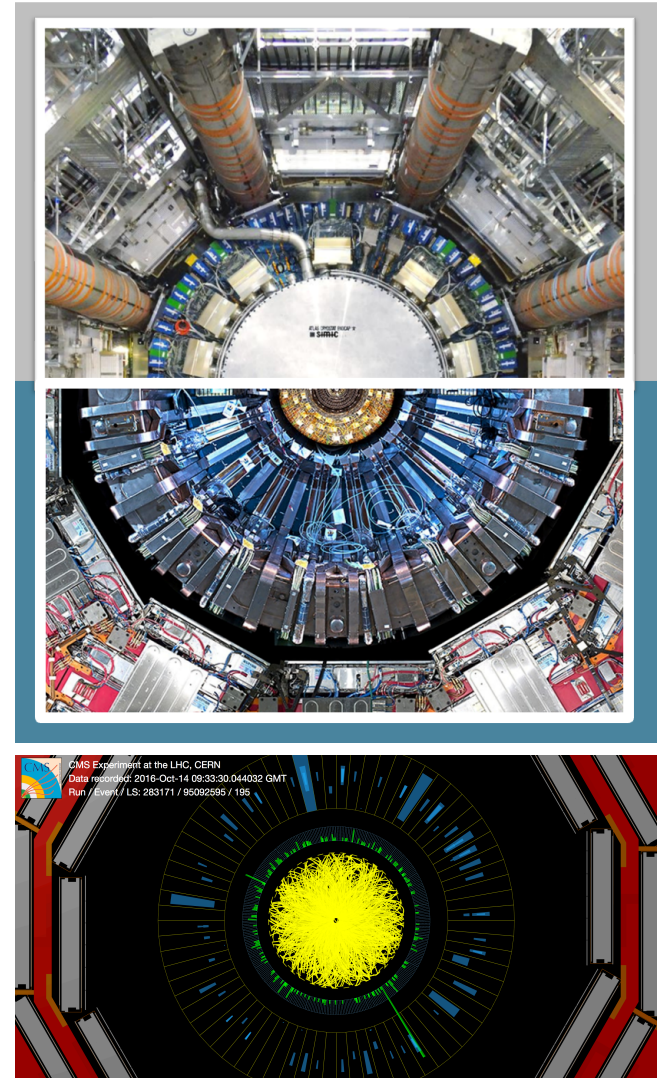
- exploit new reconstruction techniques for enhanced particle identification
- develop novel trigger concepts for **rare** or **exotic/non-standard signatures**
 - di-Higgs ($4b, b\bar{b}\tau\tau$), low-mass Higgs bosons, $HHH \rightarrow 6b$, rare B-meson decays, ...
 - long-lived particles (LLPs), LFV Higgs decays, heavy neutral leptons, heavy stable charged particles (HSCPs), soft unclustered energy patterns (SUEPs), ...
- investigate **trigger-level analysis (TLA)** in the context of Phase 2 for HL-LHC
 - reduced event sizes, storing lightweight high-level information ie. trigger objects
 - example search: di-jet resonances (best sensitivity in 0.7-1.2 TeV range in Run 3)
 - TDAQ upgrades provide new potential to explore TLA (e.g. L0-TLA)
 - assessment of EF-TLA and L0-TLA physics cases with sensitivity studies
- investigating new **trigger selections** at L0 and EF using ML for e.g. anomaly detection
- developing a trigger analysis kit to studying Phase-2 signals acceptances, rates and efficiencies



Summary and Outlook

- Next-Generation Triggers (NGT) is a cross-disciplinary, cross-experiment project to leverage innovative computing technologies for data acquisition and processing, beyond the baseline Phase-2 upgrades
- Plethora of dedicated NGT activities focusing on innovative technologies ranging from AI/ML-driven techniques to heterogenous processing on accelerators (GPUs, FPGAs, AIEs,...) for the ATLAS and CMS trigger systems
 - including scouting/TLA both at HLT/EF and LIT/L0 trigger levels
- HL-LHC will be the last hadron collider at comparable energies for decades, so we have to make the most of it by improving the trigger the best we can!

"the trigger does not decide which physics model is right, it just decides which physics model is left"



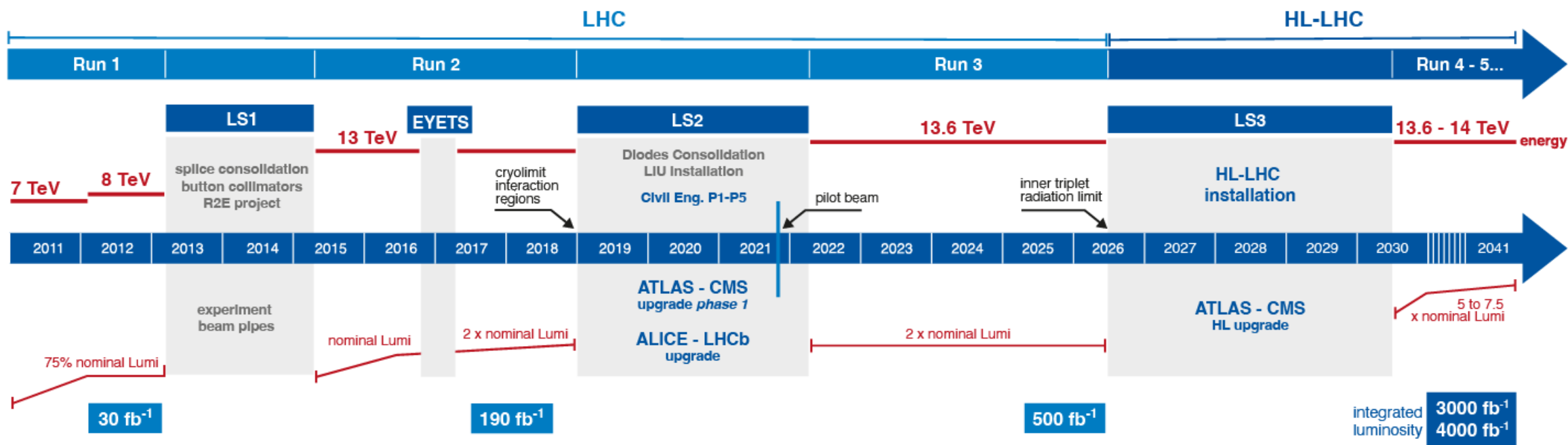
Supported by the Eric & Wendy Schmidt Fund for Strategic Innovation (grant agreement SIF-2023-004)



NextGen
Next Generation Triggers



LHC / HL-LHC Plan



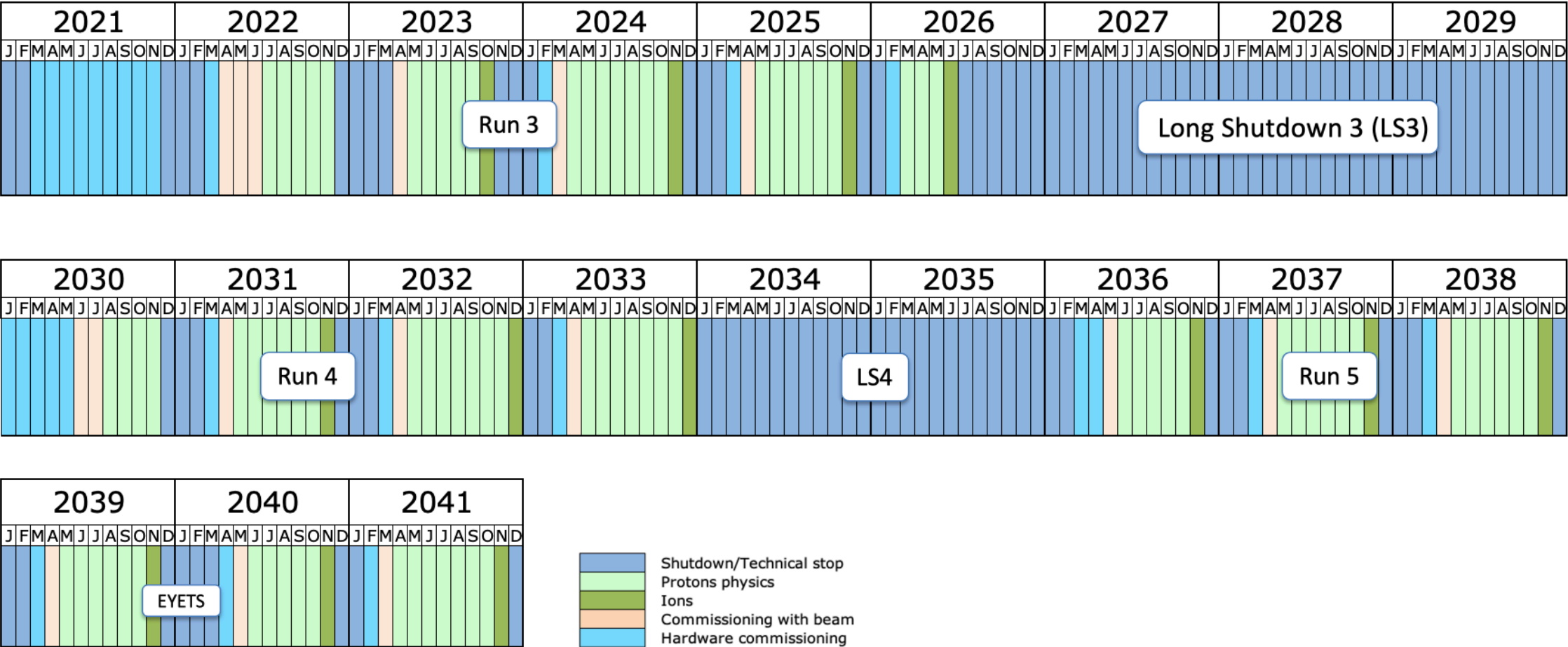
HL-LHC TECHNICAL EQUIPMENT:



HL-LHC CIVIL ENGINEERING:

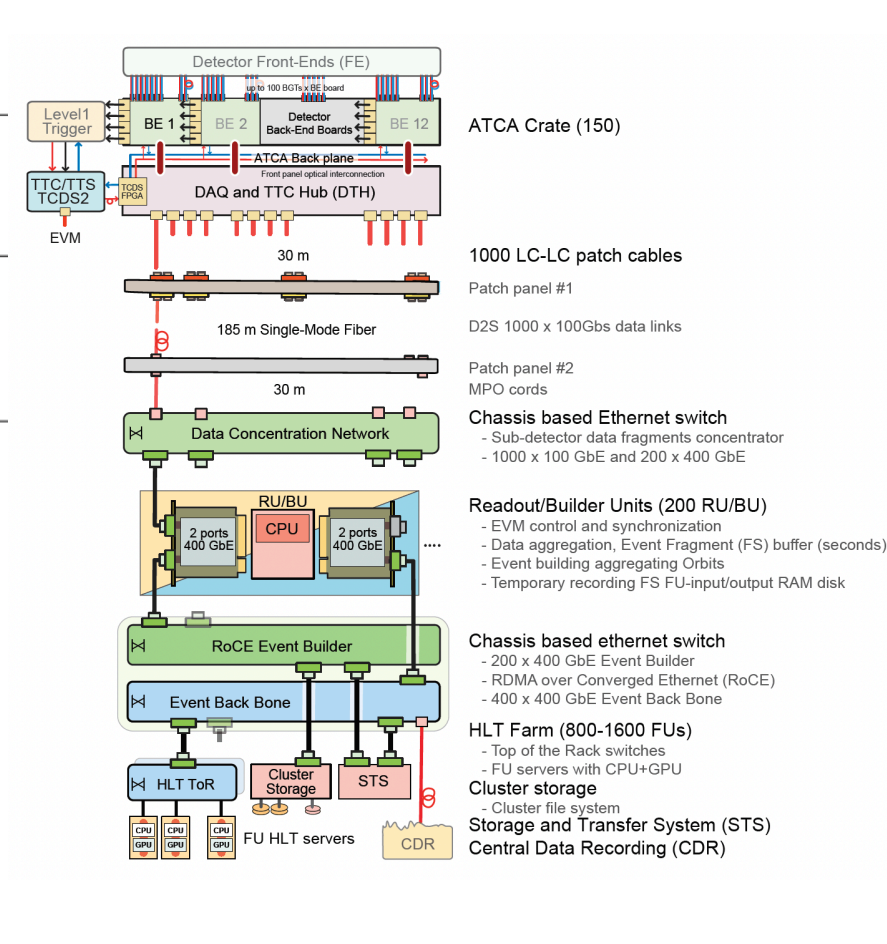
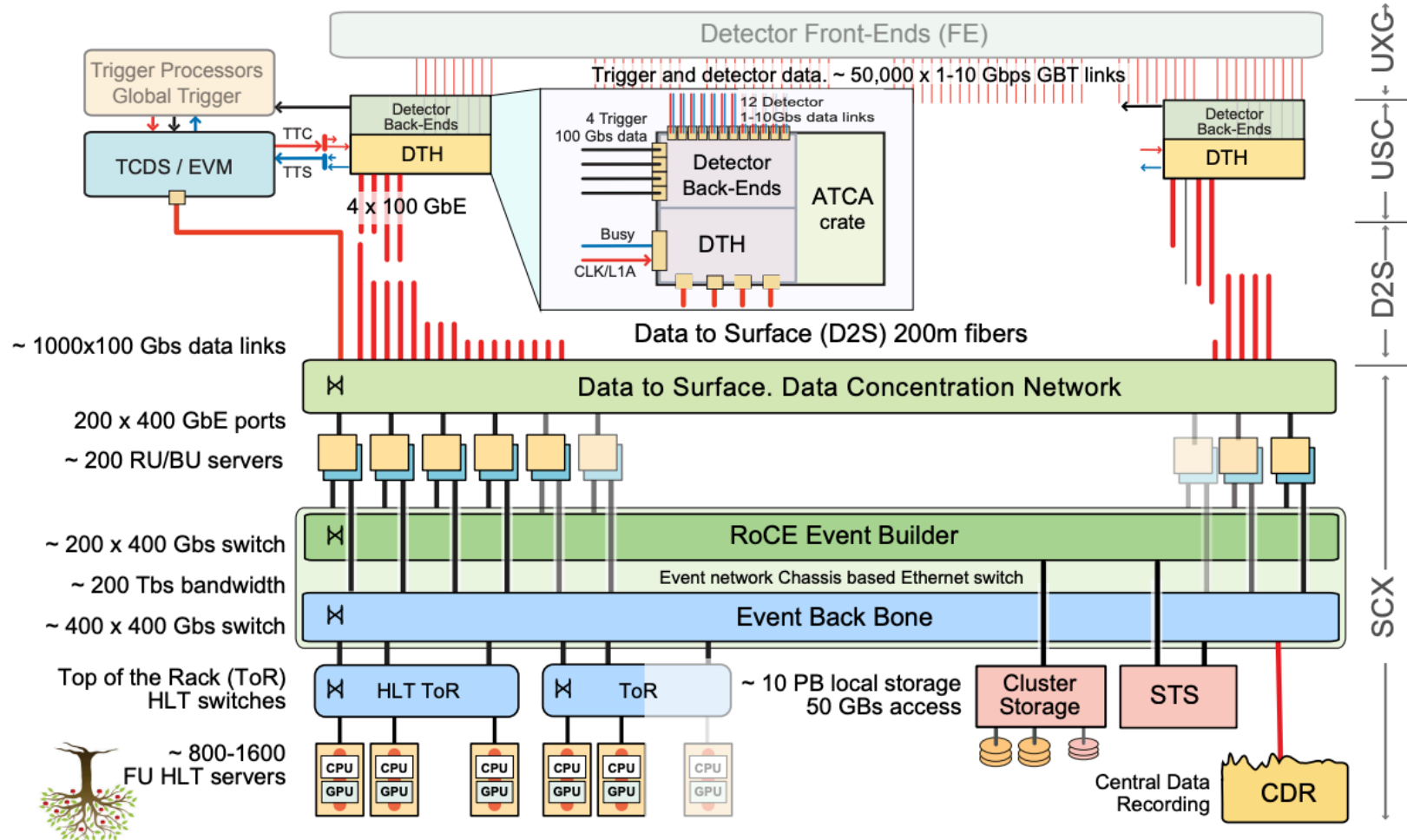


HL-LHC Schedule



Last update: November 24

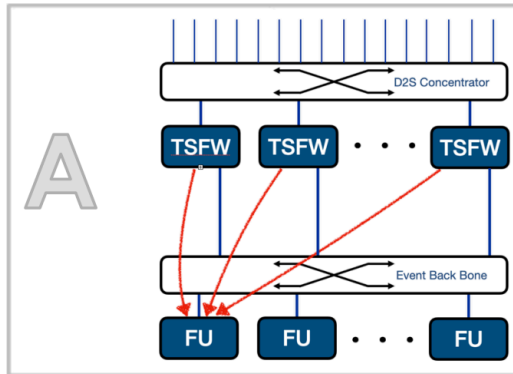
CMS Phase 2 DAQ System (TDR)



CMS Phase 2 DAQ System

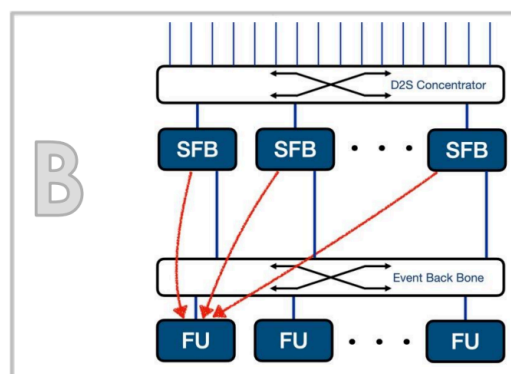
- Event Builder options discussed in [DAQ@LHC workshop](#)
 - e.g. RU/BU (DAQ3) or alternative BU/FU (à la DAQ1) → Option C:

Fragment file-based Builder



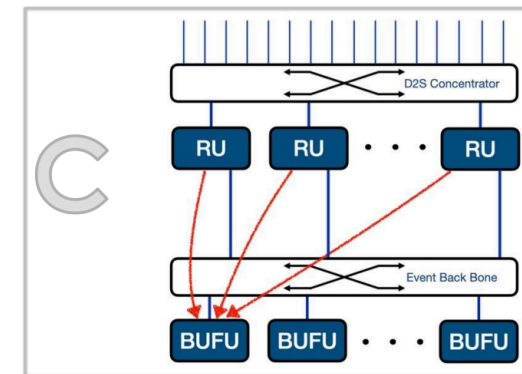
- TCP Stream File Writer (TSFW):
 - Write in a RAM disk the TCP streams from ~850 D2S Channels over 200 RU machines
- HLT CMSSW Data Input (FU):
 - Read from 200 NFS mounted RAM disks ~850 files for same orbit range
 - Parse full orbit and assemble singular events to be processed

Super-Fragment File-based Builder



- Super-Fragment builder (SFB):
 - Write in a RAM disk the Orbit Super-Fragments (grouping of D2S Channels) in each RU machine from ~850 D2S Channels
- HLT CMSSW Data Input:
 - Read from 200 NFS mounted RAM disks ~200 files for same orbit range
 - Parse full orbit and assemble singular events to be processed

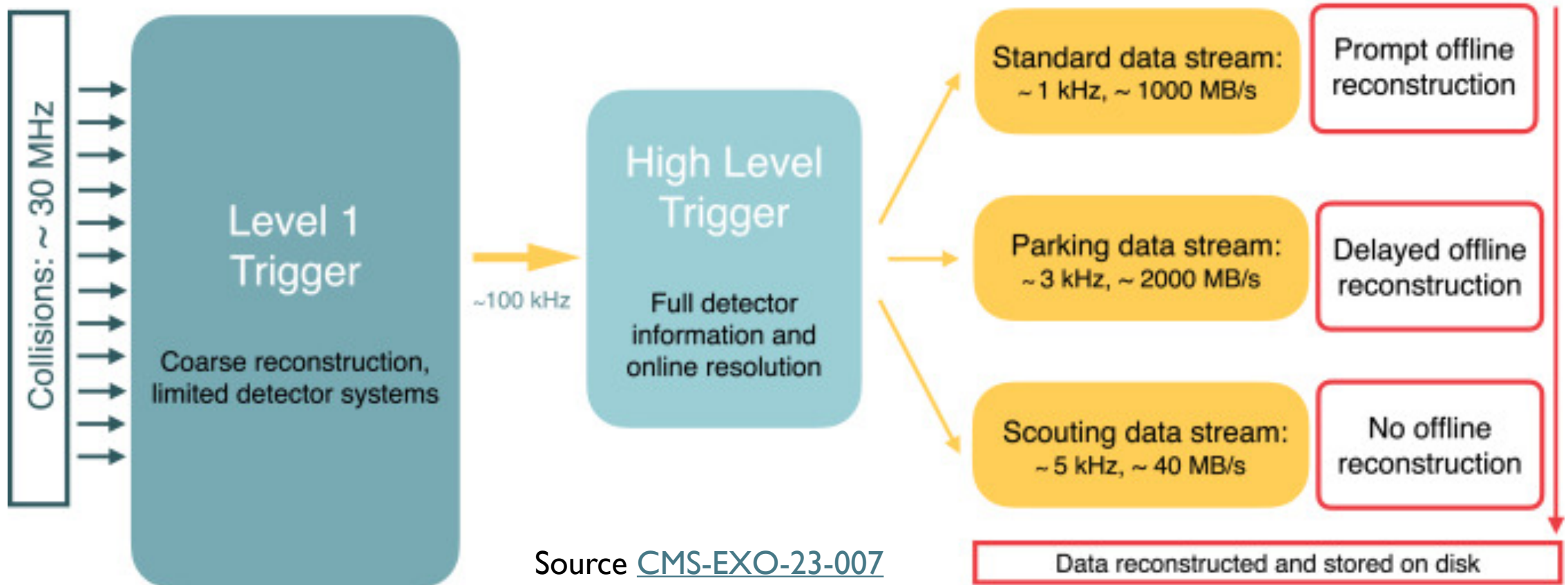
Builder running on FU



- Readout Unit (RU) builds Orbit Super-Fragment on RU machines
- Build Unit (BU) builds the Orbit in the RAM disk of FU machine
- HLT CMSSW Data Input (FU):
 - Read full orbits from local RAM disks
 - Assemble singular events to be processed

Typical Data Flow (CMS)

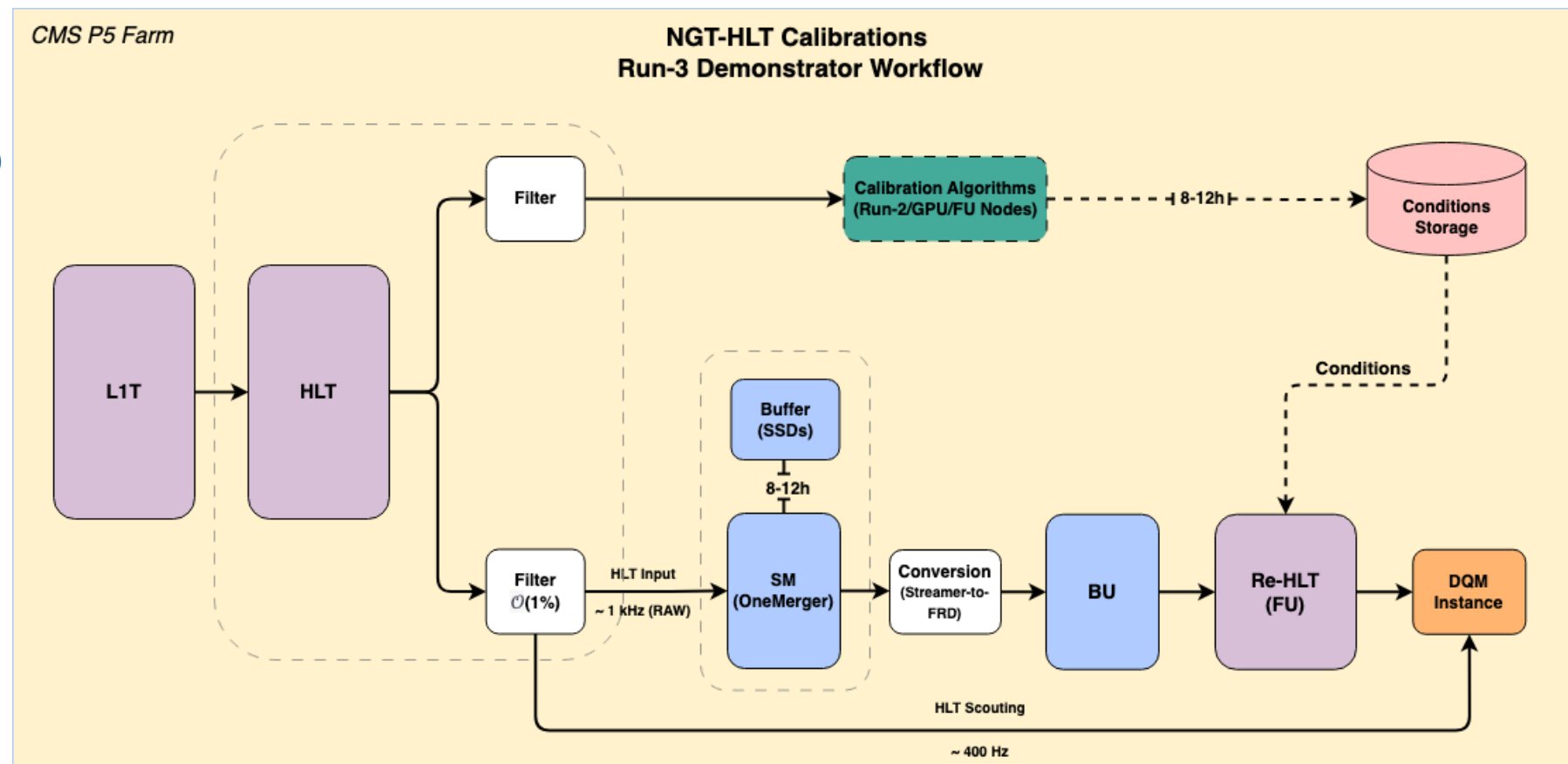
Data flow for a typical 2018 data-taking scenario



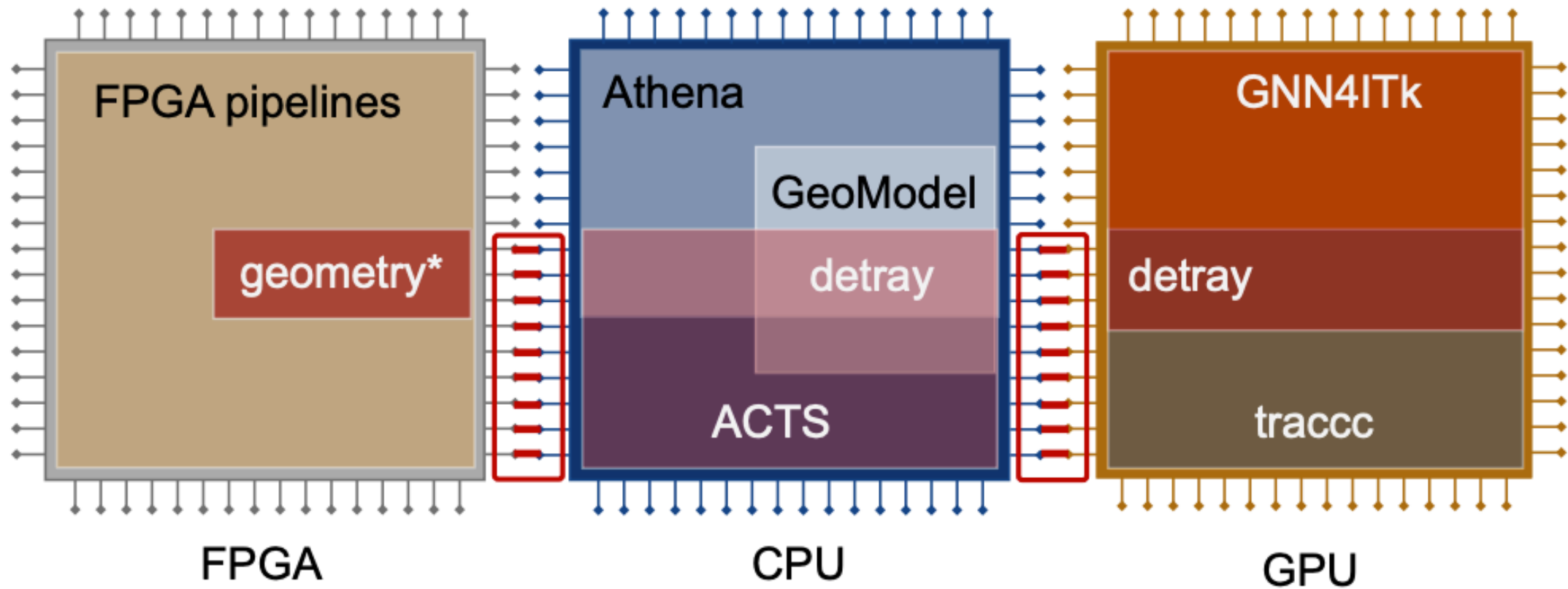
CMS (Task 3.4): Optimal Calibrations

NGT demonstrator workflow
in the DAQ system:

- buffer $\mathcal{O}(1\%)$ of HLT input data (RAW)
 - ~ 1 kHz for 8 hours
 - SSDs mounted on SM node
- derive improved calibrations
- re-run the HLT-Scouting (Re-HLT)
- store the scouting results
 - compare with the original HLT-Scouting (à la DQM)



ATLAS (Task 2.6): Common Tracking Event Filter



ATLAS (Task 2.6): Common Tracking Event Filter

