

# AI-Engines for fixed latency environments

- Speaker: **I. Xiotidis** on behalf of ATLAS TDAQ Collaboration
- Conference: FastML - Zurich
- Date: Sep-2025



**NexTGen**  
Next Generation Triggers

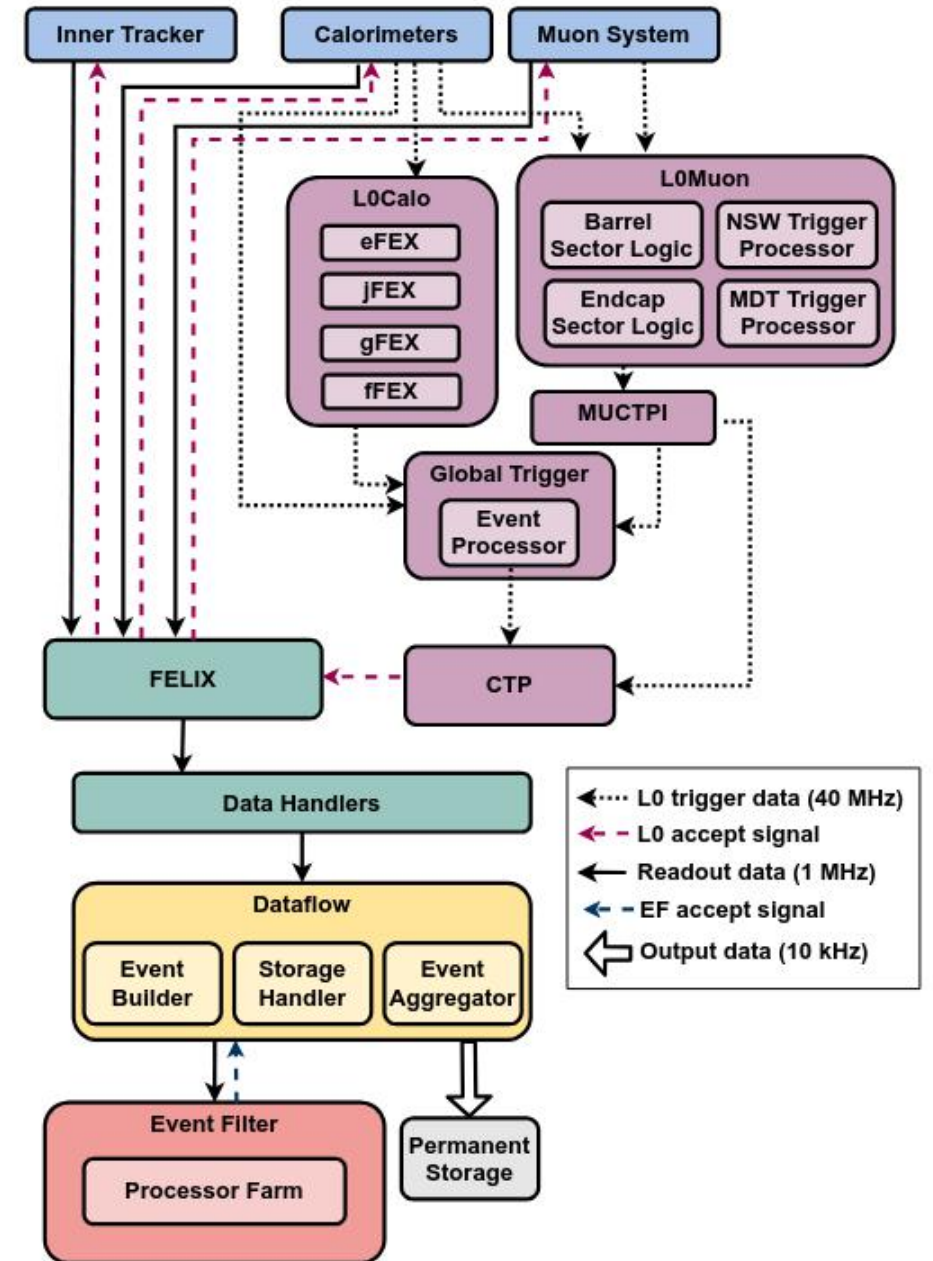
# Overview

---

- Introduction
  - Phase-II TDAQ System
  - The ATLAS Global Trigger
  - Next Generation Triggers
  - WP2.1
- AMD AI-Engines
  - Brief introduction
- Application of AIEs in fixed latency
  - Boosted Decision Tree
  - Convolutional Neural Networks
- Conclusions

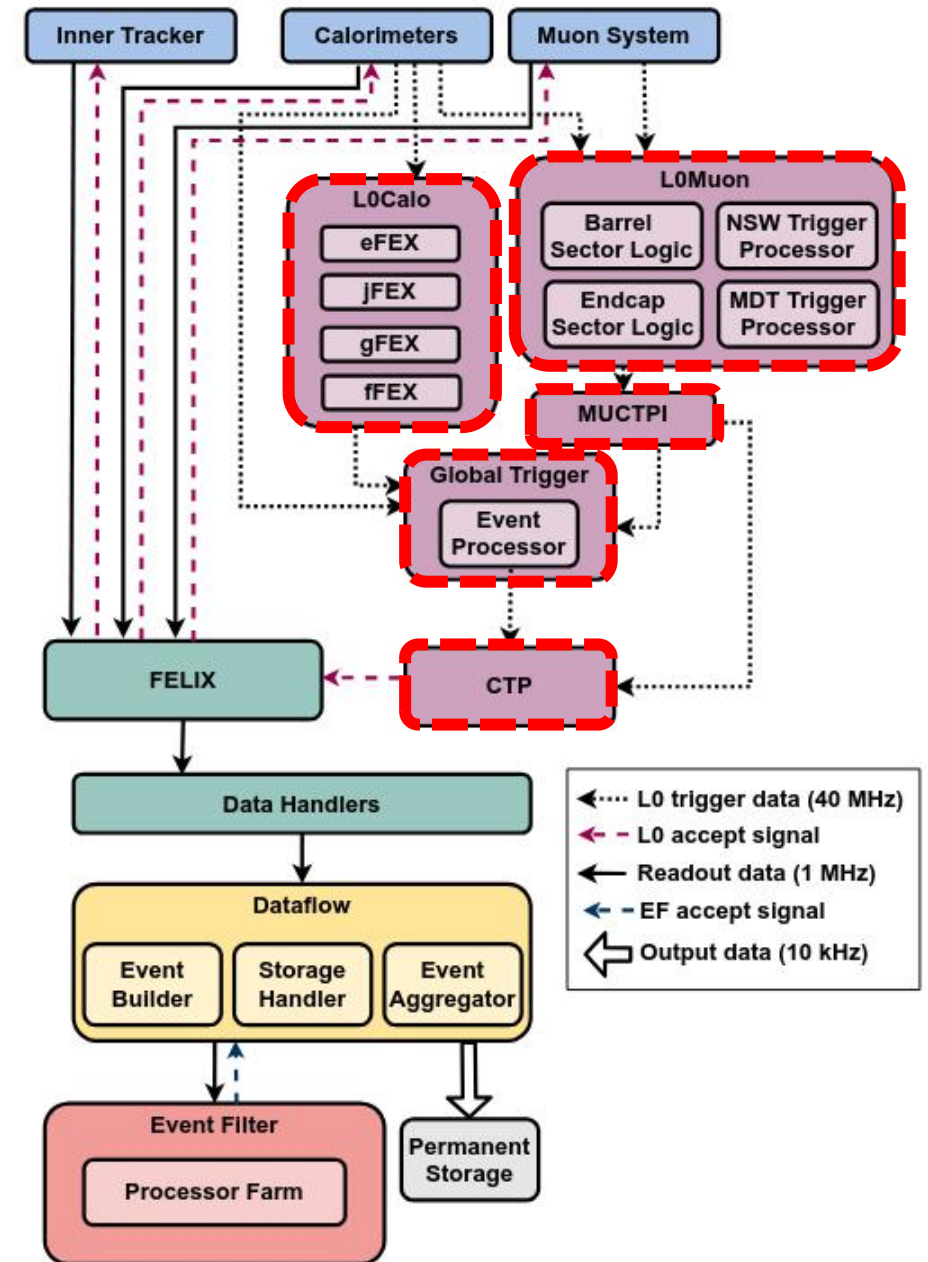
# The ATLAS Phase-II TDAQ

- For HL-LHC ATLAS will upgrade the full TDAQ system
- The system will follow the existing two stage technology with updated trigger rates and latency:



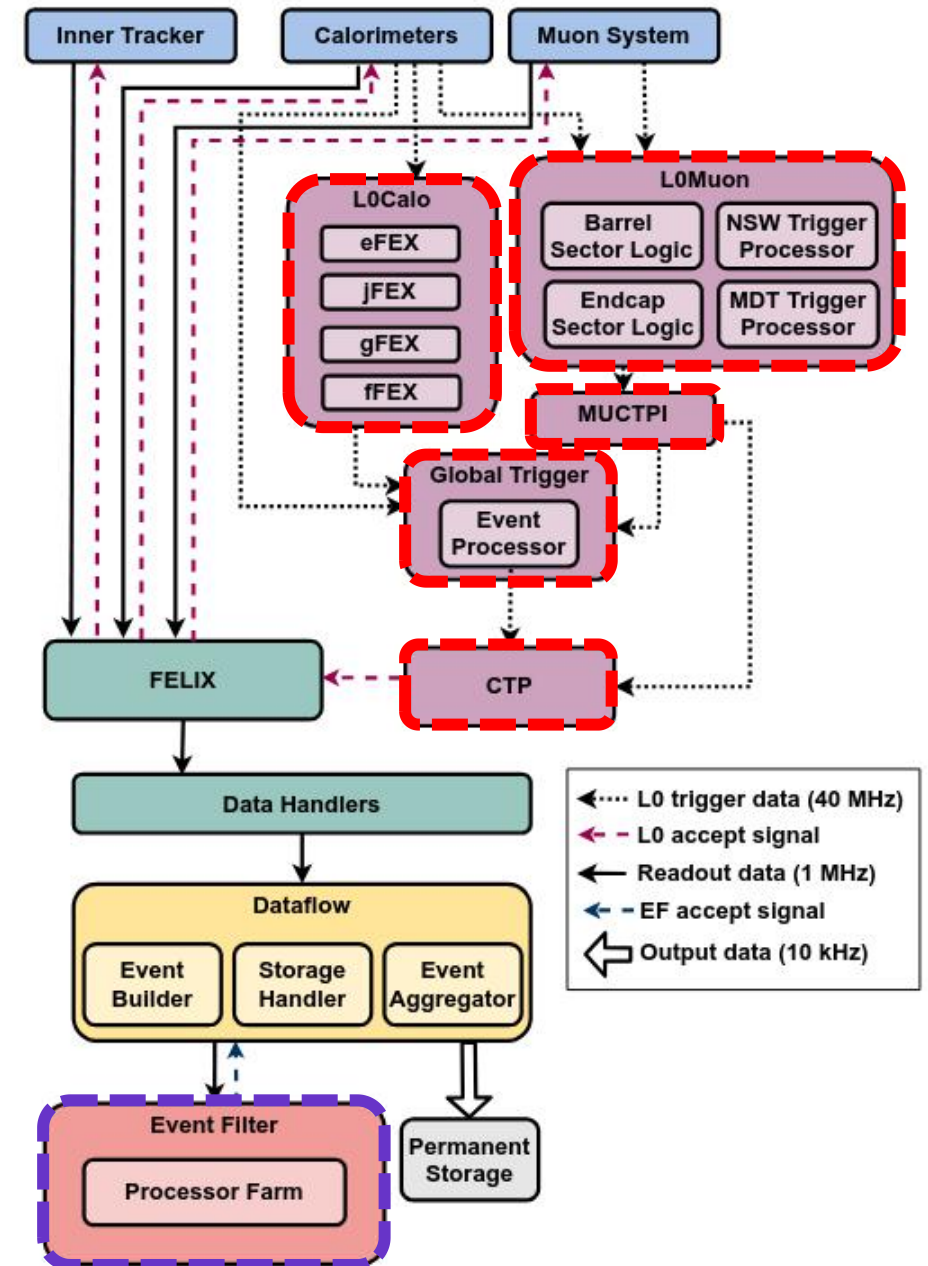
# The ATLAS Phase-II TDAQ

- For HL-LHC ATLAS will upgrade the full TDAQ system
- The system will follow the existing two stage technology with updated trigger rates and latency:
  - **Level-0**: FPGA based custom hardware
    - Input: 40MHz
    - Output: 1MHz
    - Latency: 10us



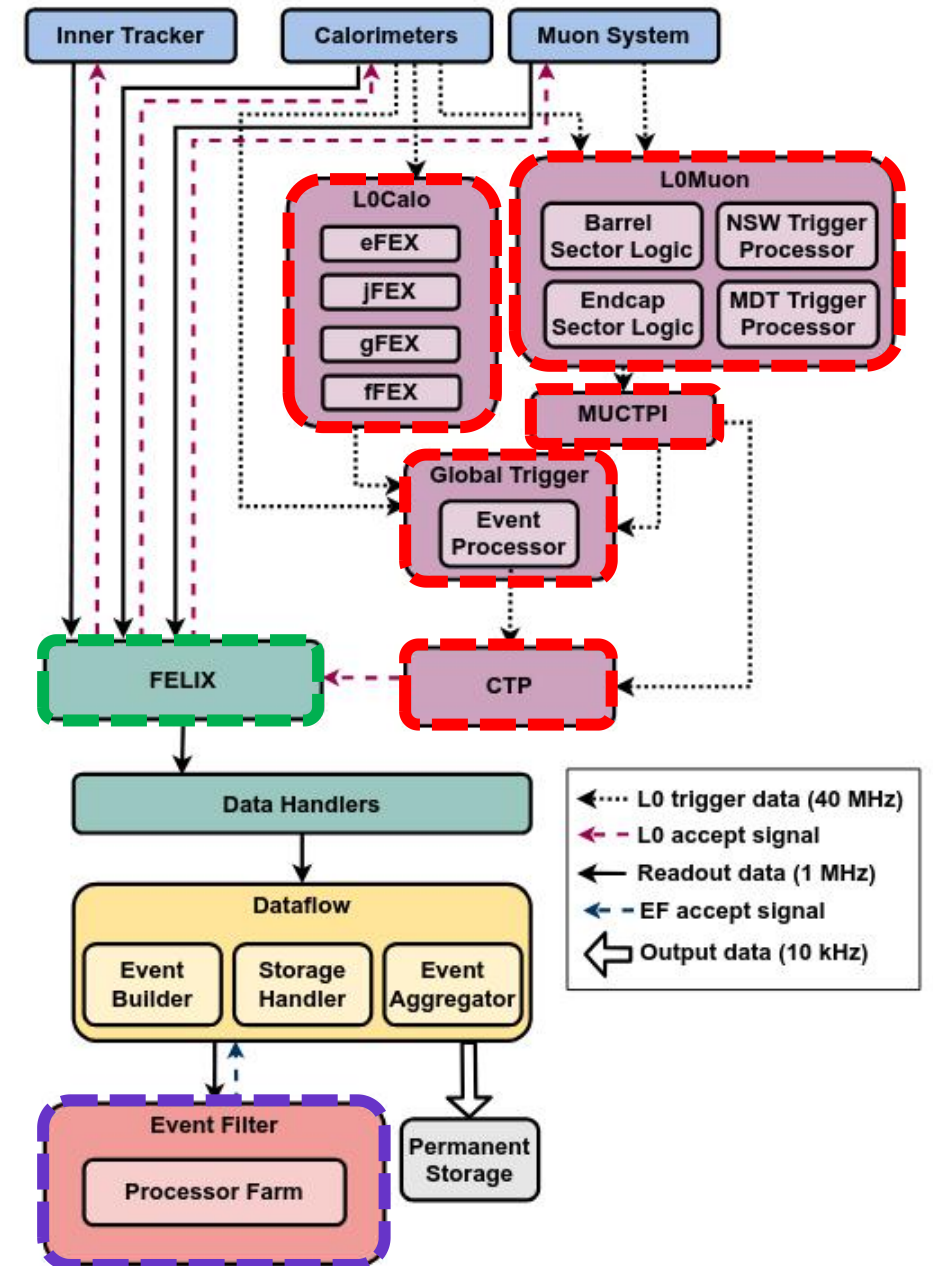
# The ATLAS Phase-II TDAQ

- For HL-LHC ATLAS will upgrade the full TDAQ system
- The system will follow the existing two stage technology with updated trigger rates and latency:
  - **Level-0**: FPGA based custom hardware
    - Input: 40MHz
    - Output: 1MHz
    - Latency: 10us
  - **Event Filter**: CPU based server farm
    - Potentially including GPUs and/or FPGA accelerators
    - Input: 1MHz from L0
    - Output: 10kHz



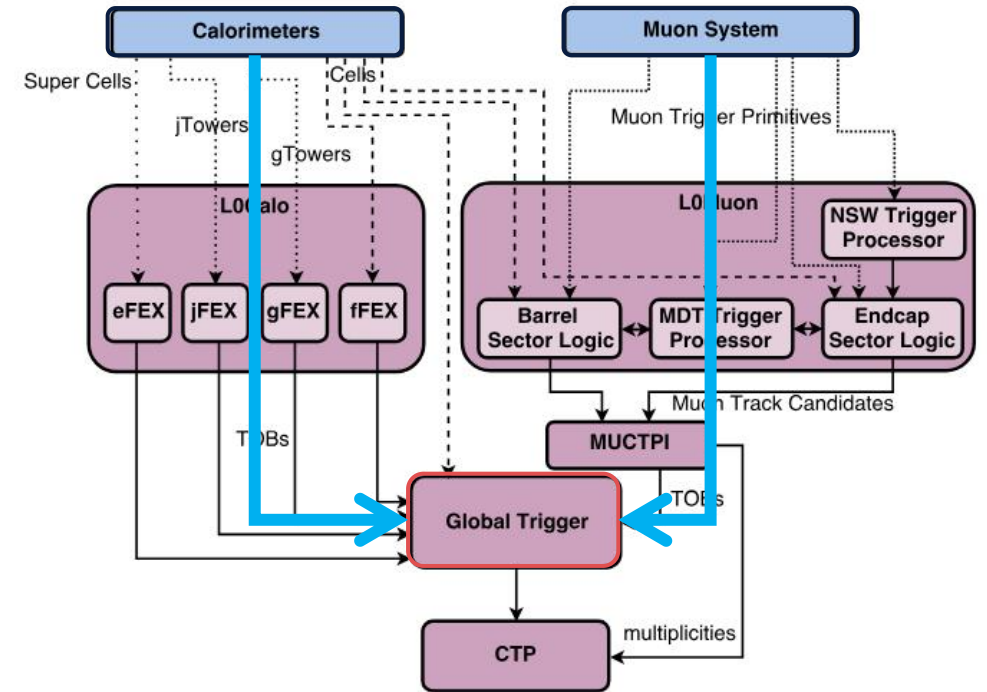
# The ATLAS Phase-II TDAQ

- For HL-LHC ATLAS will upgrade the full TDAQ system
- The system will follow the existing two stage technology with updated trigger rates and latency:
  - **Level-0**: FPGA based custom hardware
    - Input: 40MHz
    - Output: 1MHz
    - Latency: 10us
  - **Event Filter**: CPU based server farm
    - Potentially including GPUs and/or FPGA accelerators
    - Input: 1MHz from L0
    - Output: 10kHz
- Data handling and routing all via **FELIX** cards for minimum latency



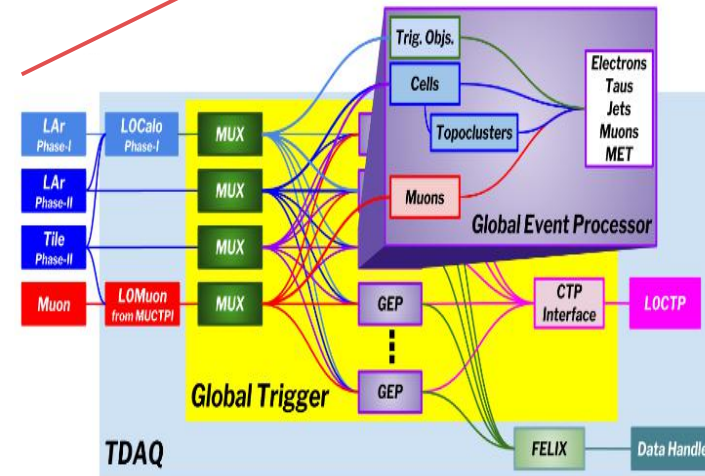
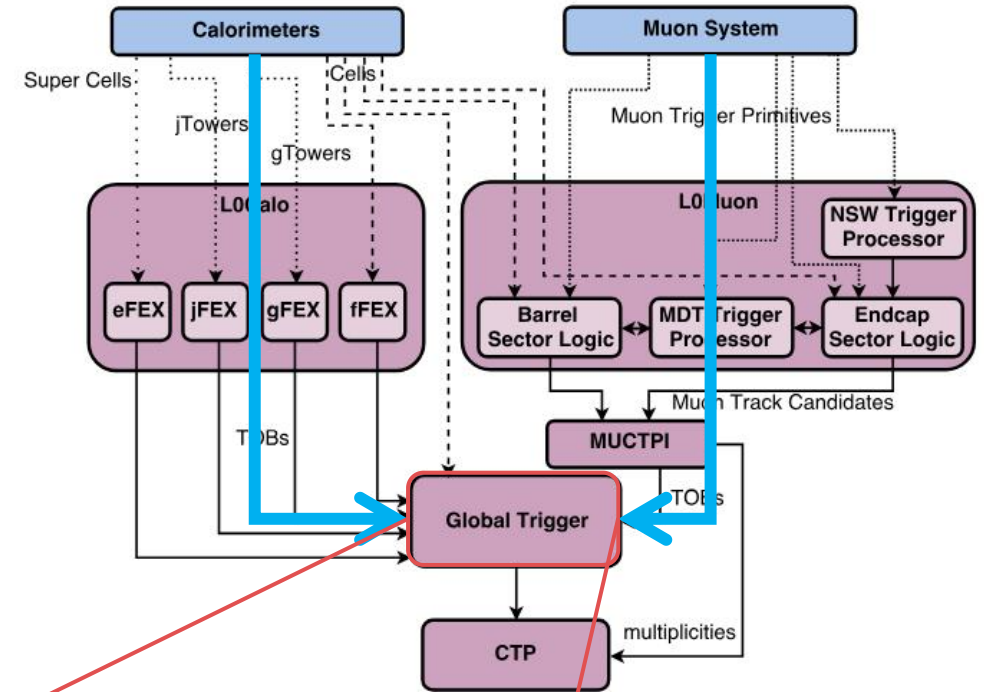
# The ATLAS Global Trigger

- Receives data from the **calorimeter** and **muon** sub-systems
  - 50Tbps of data streamed through
  - Processes all the data in  $\sim 10\mu\text{s}$  latency



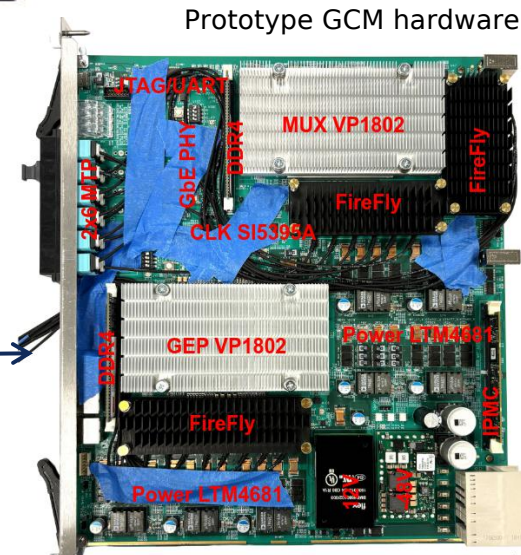
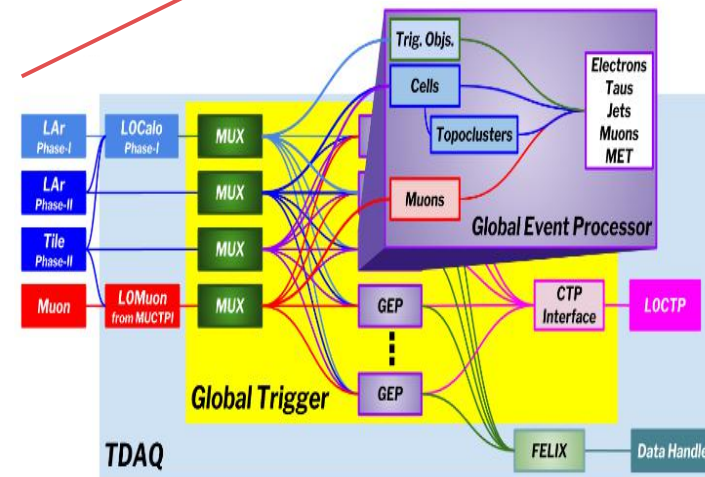
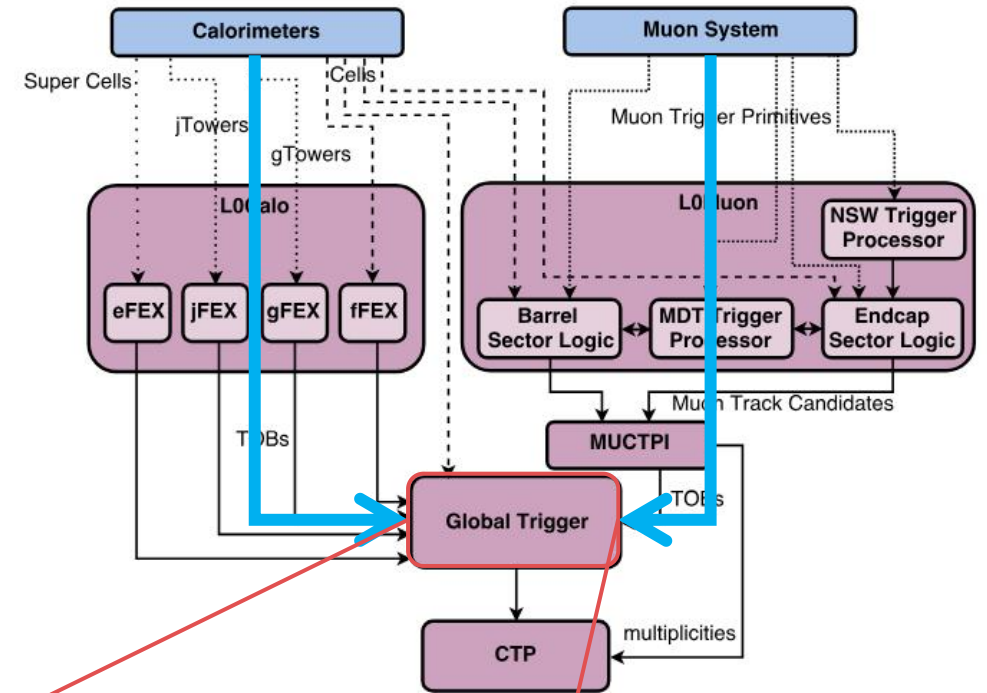
# The ATLAS Global Trigger

- Receives data from the **calorimeter** and **muon** sub-systems
  - 50Tbps of data streamed through
  - Processes all the data in ~10us latency
- The **Global Trigger** executes offline-like trigger algorithms
  - Uses the full granularity information of the ATLAS calorimeter
  - Associates calorimeter objects with muons for more efficient triggering



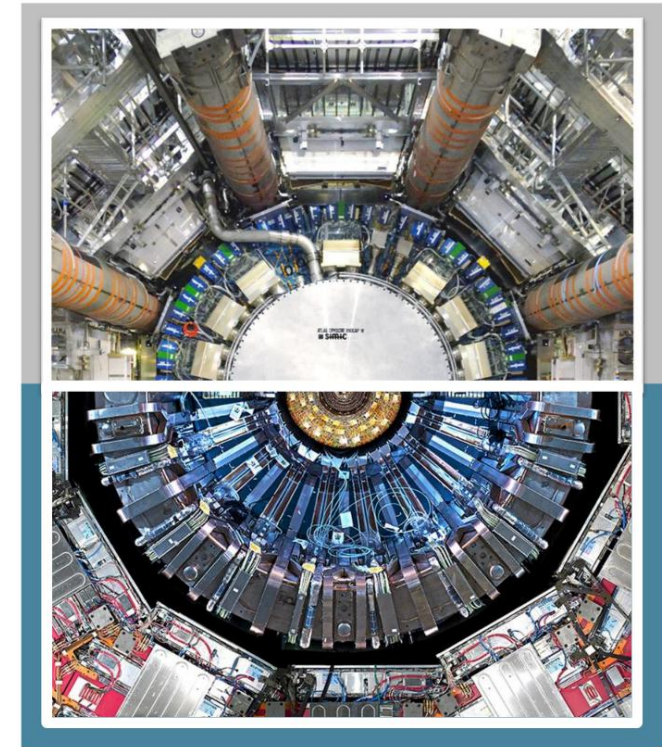
# The ATLAS Global Trigger

- Receives data from the **calorimeter** and **muon** sub-systems
  - 50Tbps of data streamed through
  - Processes all the data in  $\sim 10\mu\text{s}$  latency
- The **Global Trigger** executes offline-like trigger algorithms
  - Uses the full granularity information of the ATLAS calorimeter
  - Associates calorimeter objects with muons for more efficient triggering
- To handle the large amount of input links and offline-like decision mechanisms
  - The events are time-multiplexed requiring  $O(50)$  boards for the full system
  - Large FPGA is required (AMD Versal Premium 1802) with many I/O links and resources



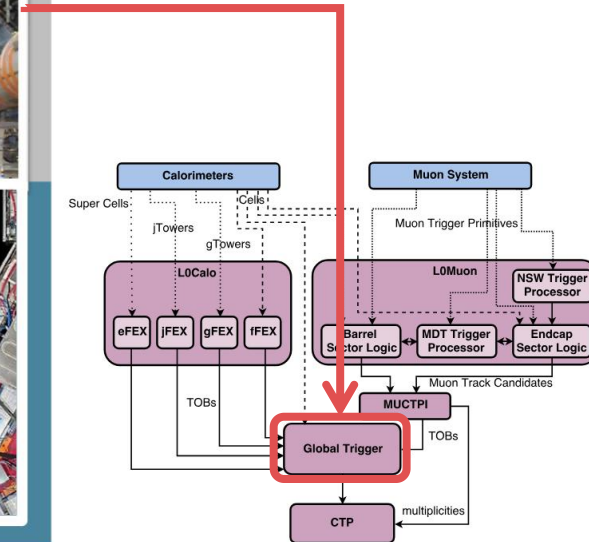
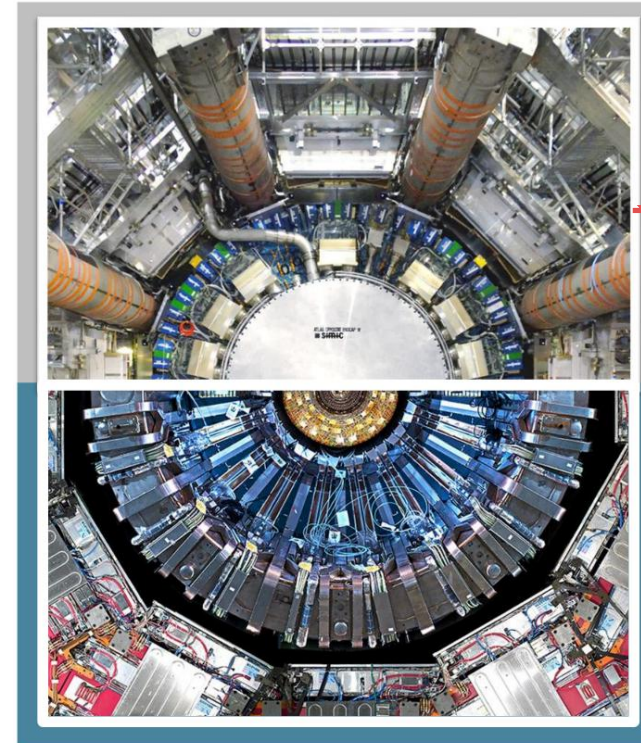
# NextGeneration Trigger project

- The NextGeneration Trigger project is an R&D activity within CERN covering 4x areas (WPs):
  - **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
- Funded for 4x years (2024-2028) with main goals:
  - Explore trigger algorithms beyond the Phase-II base plan
  - Identify improvements on LHC experiments beyond 2028+



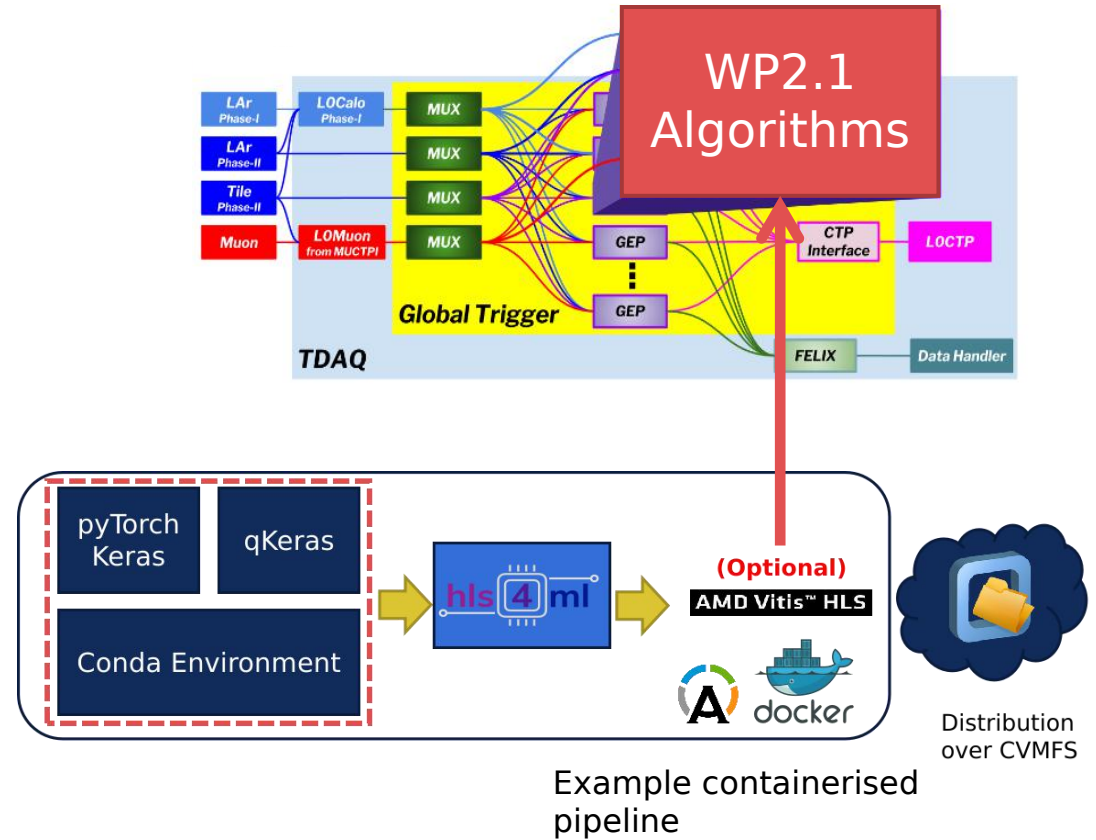
# NextGeneration Trigger project

- The NextGeneration Trigger project is an R&D activity within CERN covering 4x areas (WPs):
  - 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
- Funded for 4x years (2024-2028) with main goals:
  - Explore trigger algorithms beyond the Phase-II base plan
  - Identify improvements on LHC experiments beyond 2028+
- ATLAS has 7(!) sub work-packages covering most of TDAQ
  - WP2.1: Calorimeter based hardware triggers**



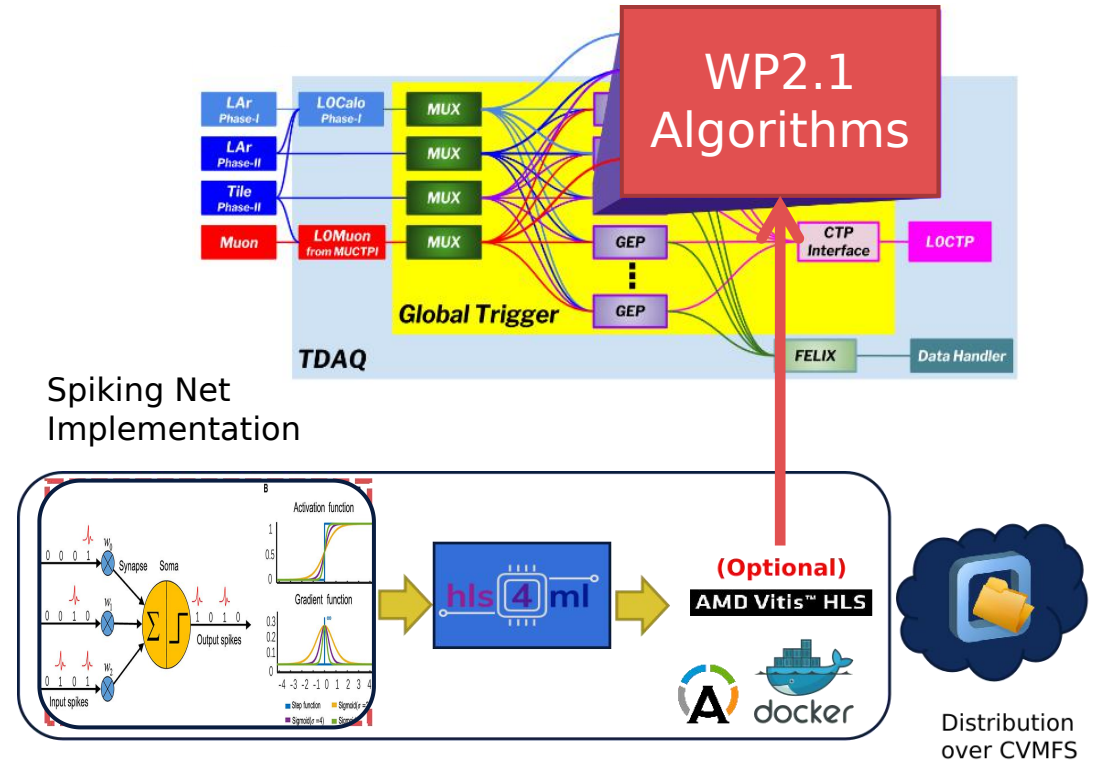
# The NextGen WP2.1

- **Goal:** Exploration of novel Machine Learning solutions for the Global Trigger
  - Development of common software framework for ML deployment within the Global ecosystem



# The NextGen WP2.1

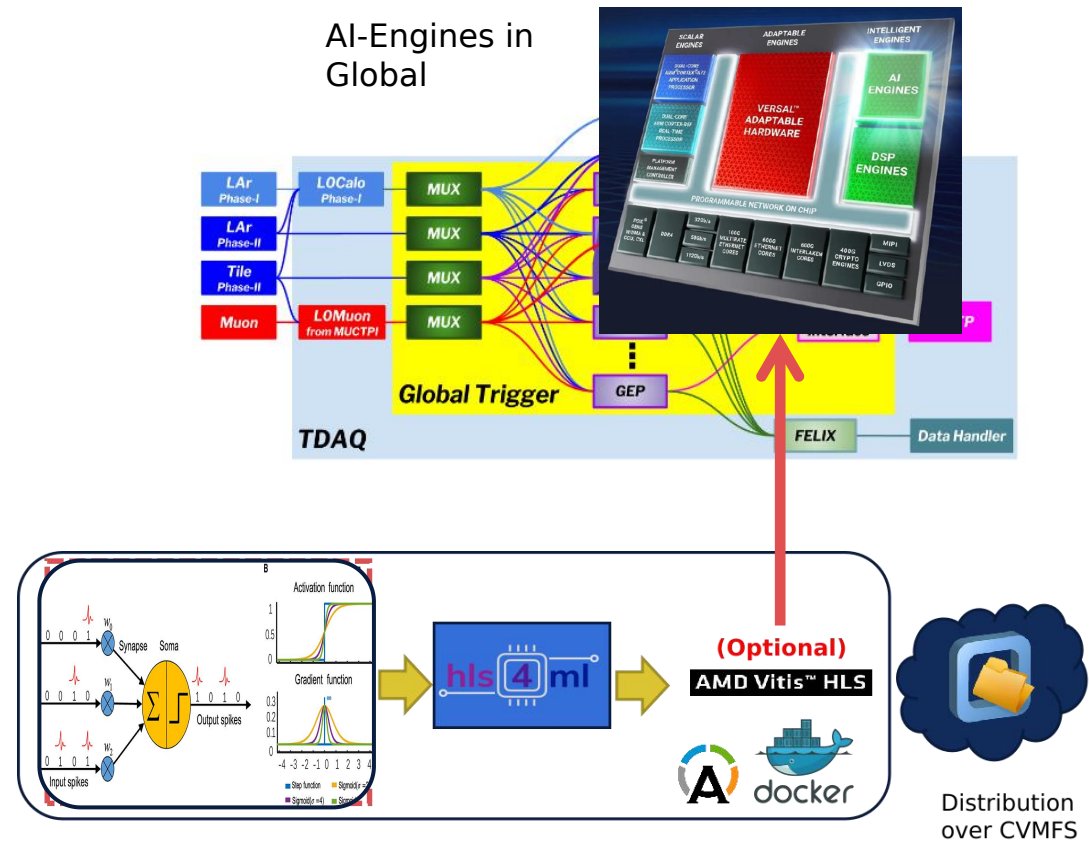
- **Goal:** Exploration of novel Machine Learning solutions for the Global Trigger
  - Development of common software framework for ML deployment within the Global ecosystem
  - Evaluation of novel ML model architectures (i.e. SNN) for multi-dimensional optimisation (power, latency, resources, etc.)



# The NextGen WP2.1

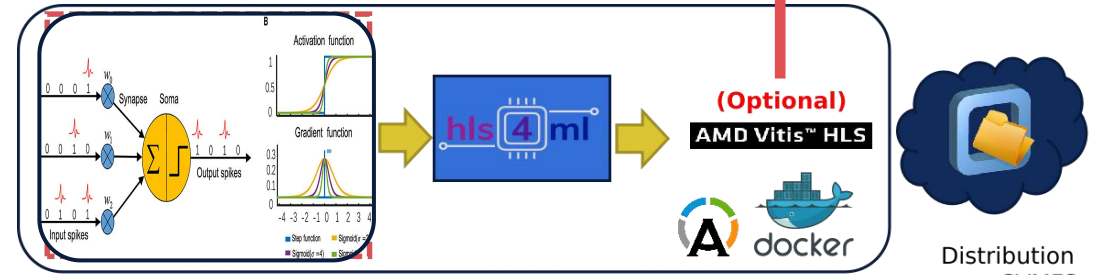
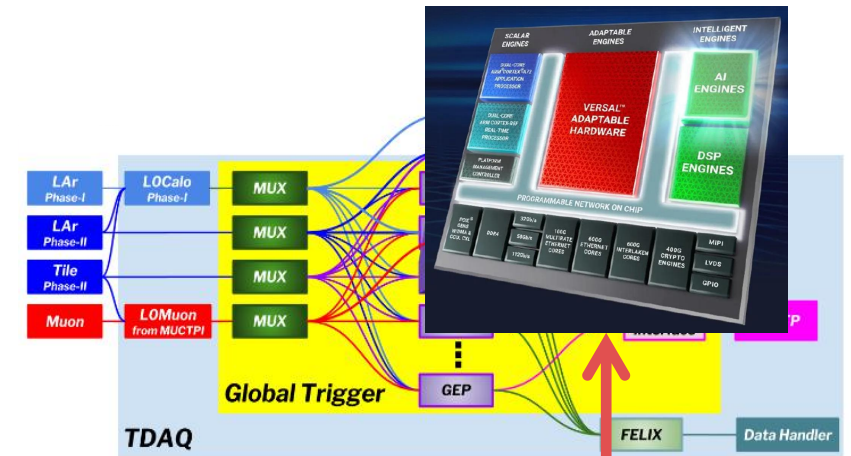
- **Goal:** Exploration of novel Machine Learning solutions for the Global Trigger
  - Development of common software framework for ML deployment within the Global ecosystem
  - Evaluation of novel ML model architectures (i.e. SNN) for multi-dimensional optimisation (power, latency, resources, etc.)
  - Exploration of new industry solutions for FPGAs

AI-Engines in Global



# The NextGen WP2.1

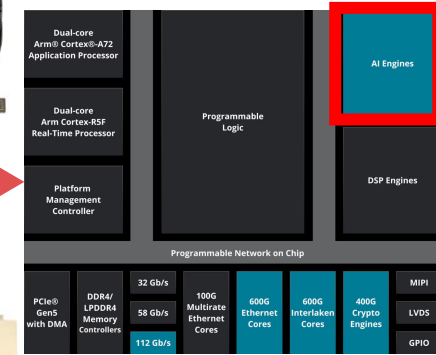
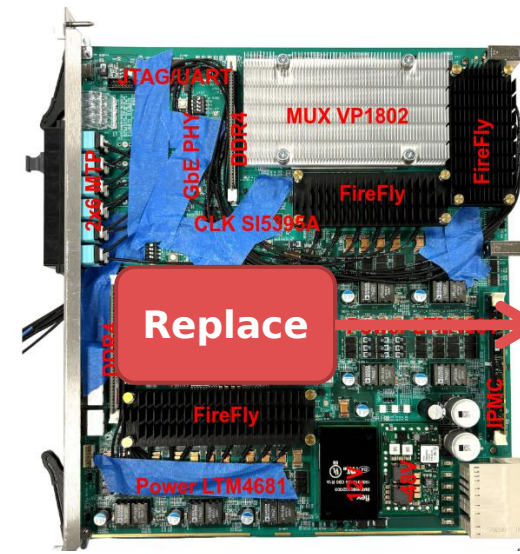
- **Goal:** Exploration of novel Machine Learning solutions for the Global Trigger
  - Development of common software framework for ML deployment within the Global ecosystem
  - Evaluation of novel ML model architectures (i.e. SNN) for multi-dimensional optimisation (power, latency, resources, etc.)
  - Exploration of new industry solutions for FPGAs
- Team:
  - Pls: N. Konstantinidis (UCL), D. Miller (U. Chicago)
  - Fellows: I. Xiotidis (CERN), D. Reikher (CERN)
  - PhD: T. Du (CERN), P. Mucha (CERN), V. Petrovic (CERN)
  - **One more fellow to be hired in 2026!**



WP2.1@ CERN

# AMD AI-Engines in Global

- The final global hardware design review is by 2026
  - AMD released a pin-compatible FPGA with current target which includes the AI-Engines (VP2802)

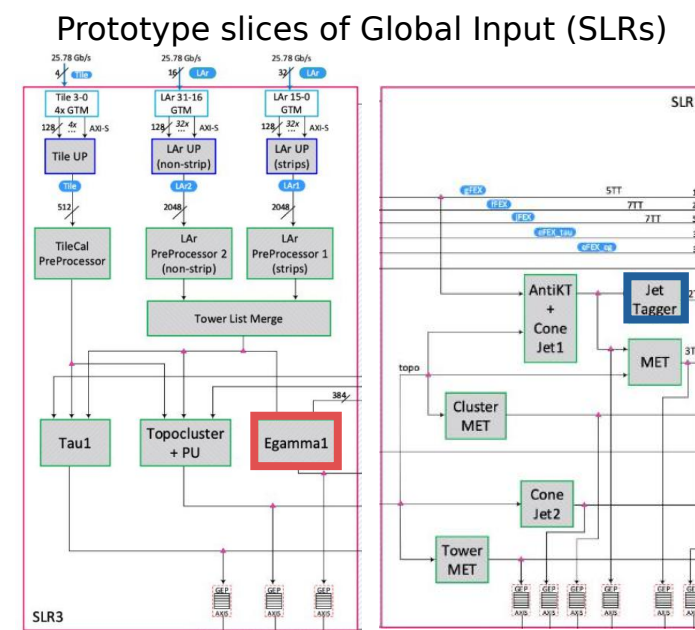






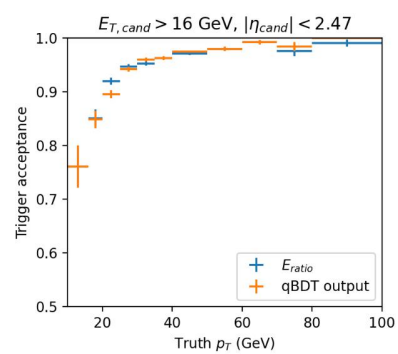
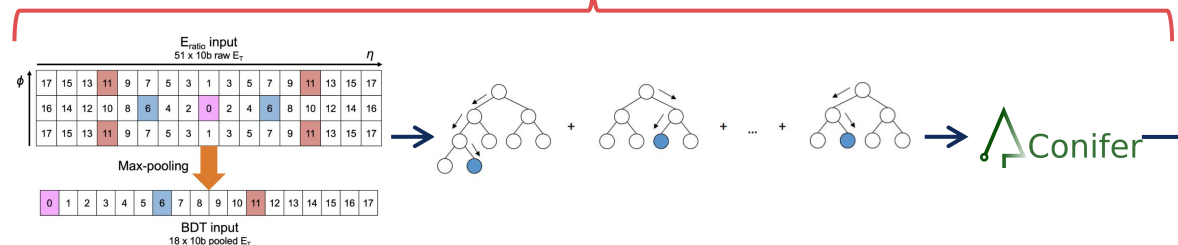
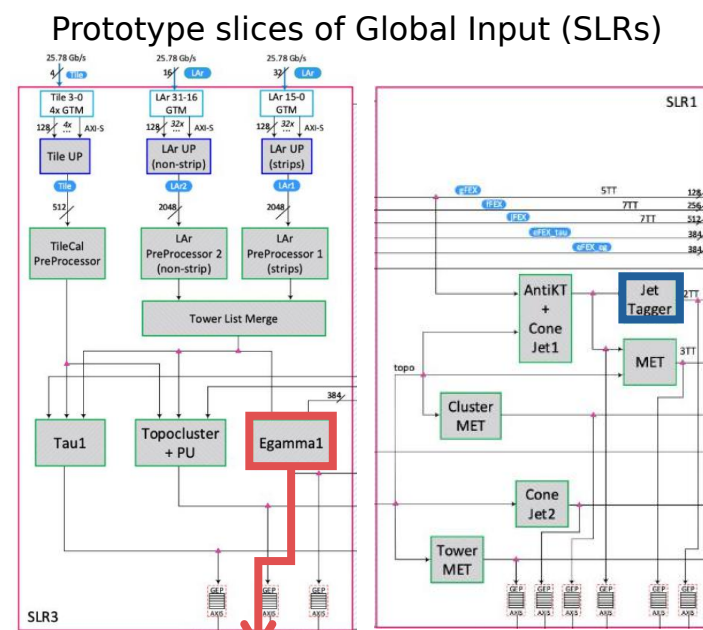
# Global models for AIEs

- Currently most models looked into global are either BDTs or CNNs
  - BDTs: **e/gamma-ID**, tau-ID
  - CNNs: Jet Seeding, Pile-Up suppression, **large-R jet tagging**



# Global models for AIEs

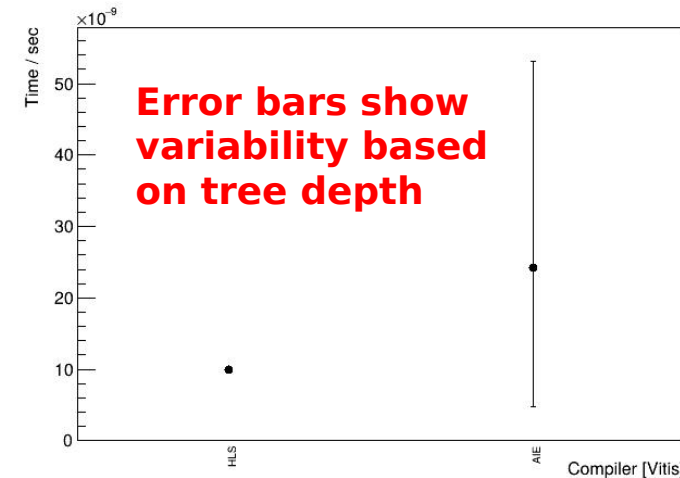
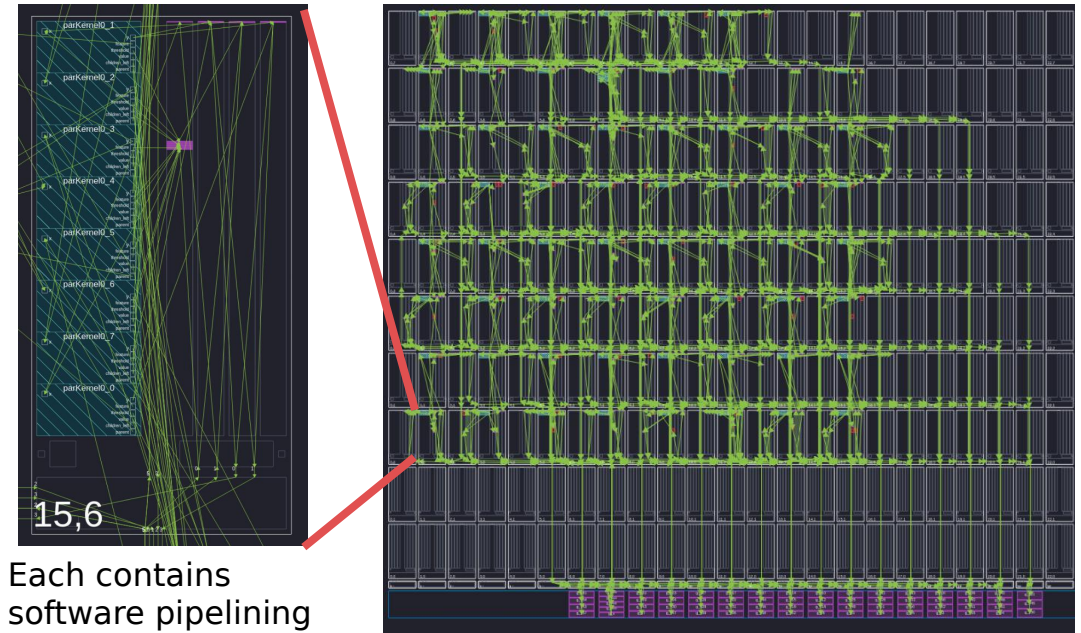
- Currently most models looked into global are either BDTs or CNNs
  - BDTs: **e/gamma-ID**, tau-ID
  - CNNs: Jet Seeding, Pile-Up suppression, **large-R jet tagging**
- First implementation for AIEs using the BDT developed for e/gamma-ID ([UCL](#))
  - Compact RTL implementation with better performance than Eratio
  - 63x Trees with fixed depth (5)



Resource	qBDT (@270MHz)
FFs	4908
BRAMs (kB)	5.5
LUTs	5376
DSP	0
Latency	3

# Global models for AIEs

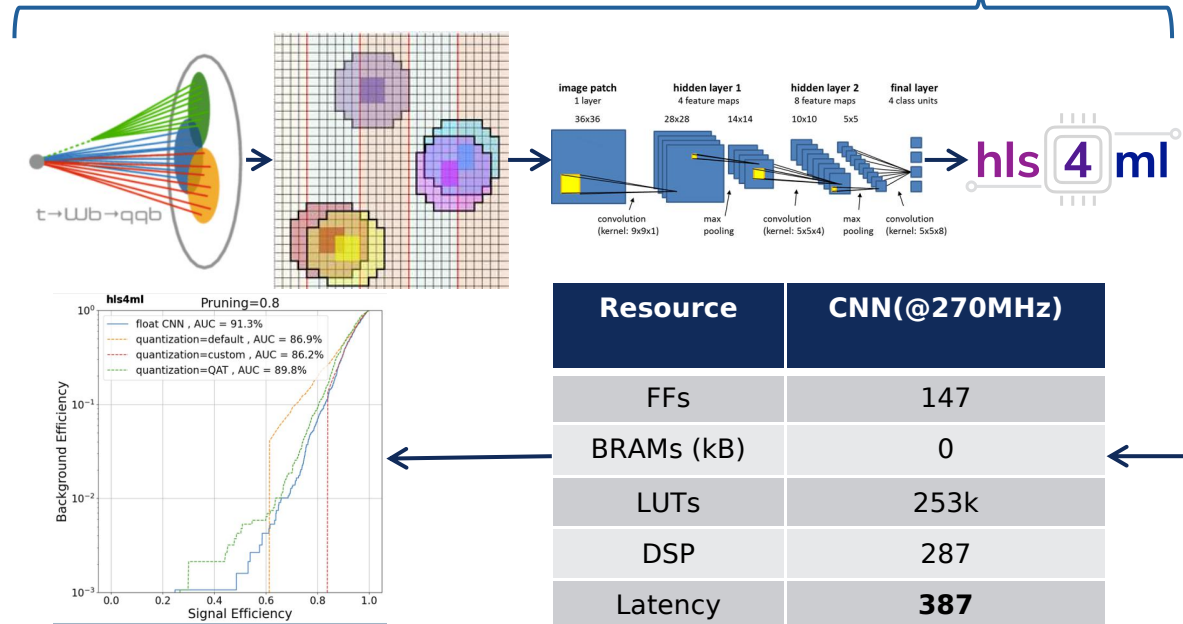
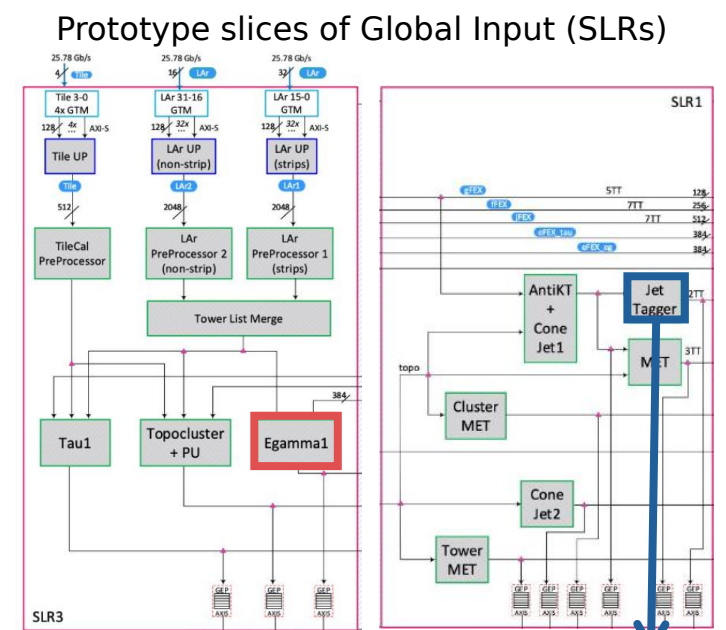
- Currently most models looked into global are either BDTs or CNNs
  - BDTs: **e/gamma-ID**, tau-ID
  - CNNs: Jet Seeding, Pile-Up suppression, **large-R jet tagging**
- First implementation for AIEs using the BDT developed for e/gamma-ID ([UCL](#))
  - Compact RTL implementation with better performance than Eratio
  - 63x Trees with fixed depth (5)
- Straightforward implementation with data pre-processing in Prog. Logic
  - Perform SIMD comparisons for the full tree depth (floating point)
  - Kernel allows for variable size tree depth
  - Tree decision in Prog. Logic



Kernel only latency no I/O

# CNN for large-R jets classification

- BDT showed promising results but computationally not that demanding!
- Second implementation on large-R jet classification for top decays ([U. Chicago](#))
  - Inspired by “[Pulling out all the tops](#)”
  - Architecture: 3x Conv2D layers, 2x Dense layers
  - #parameters: 2400





# Conclusions



- ATLAS upgrades its trigger strategy for Run-4 to cop with the HL-LHC conditions
  - Level-0 Global is an exciting upgrade opening many venues for new areas of the phase space
- The Next Generation Triggers project allows for extensive R&D on cutting edge ML and technology without limiting resources from the baseline plan
  - WP2.1 opts for enhancing the Global Trigger with Machine Learning and new industry technologies (AI-Engines)
- Heterogeneous architectures seems to be the future in data processing even within the AMD FPGAs (Versal packages)
  - AI-Engines a significant resource embedded within the package
  - Pin-compatible FPGA for Global provides a unique opportunity if the physics gain is there
- The e/gamma BDT and the large-R jet classifier CNN showed that with 32-bit precision similar latencies can be achieved with AIEs
  - Further studies are required to integrate kernel code into simulation framework and hardware validation
- Stay tuned for more WP2.1 achievements!





**NextGen**  
Next Generation Triggers