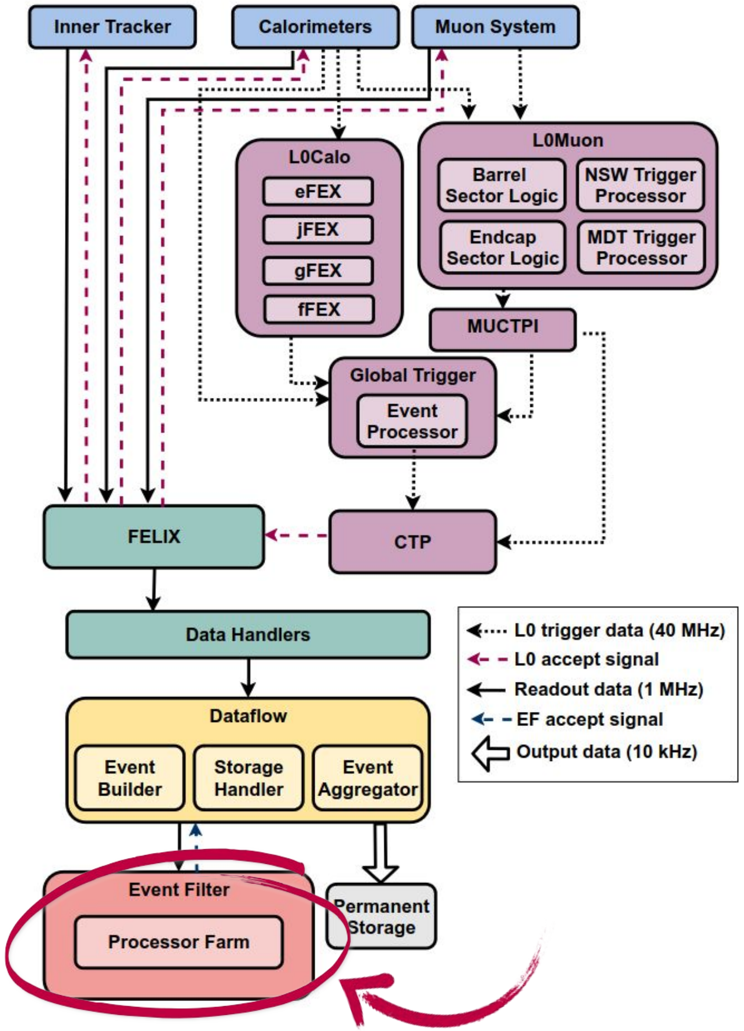
WP 2: Enhancing the ATLAS Trigger and Data Acquisition

Task 2.4: Event Filter Tracking

Noemi Calace, Stephanie Majewski
Pierfrancesco Butti, Benjamin Huth, Julian Wollrath

Next Generation Triggers – 1st Technical Workshop
25-26-27 November 2024

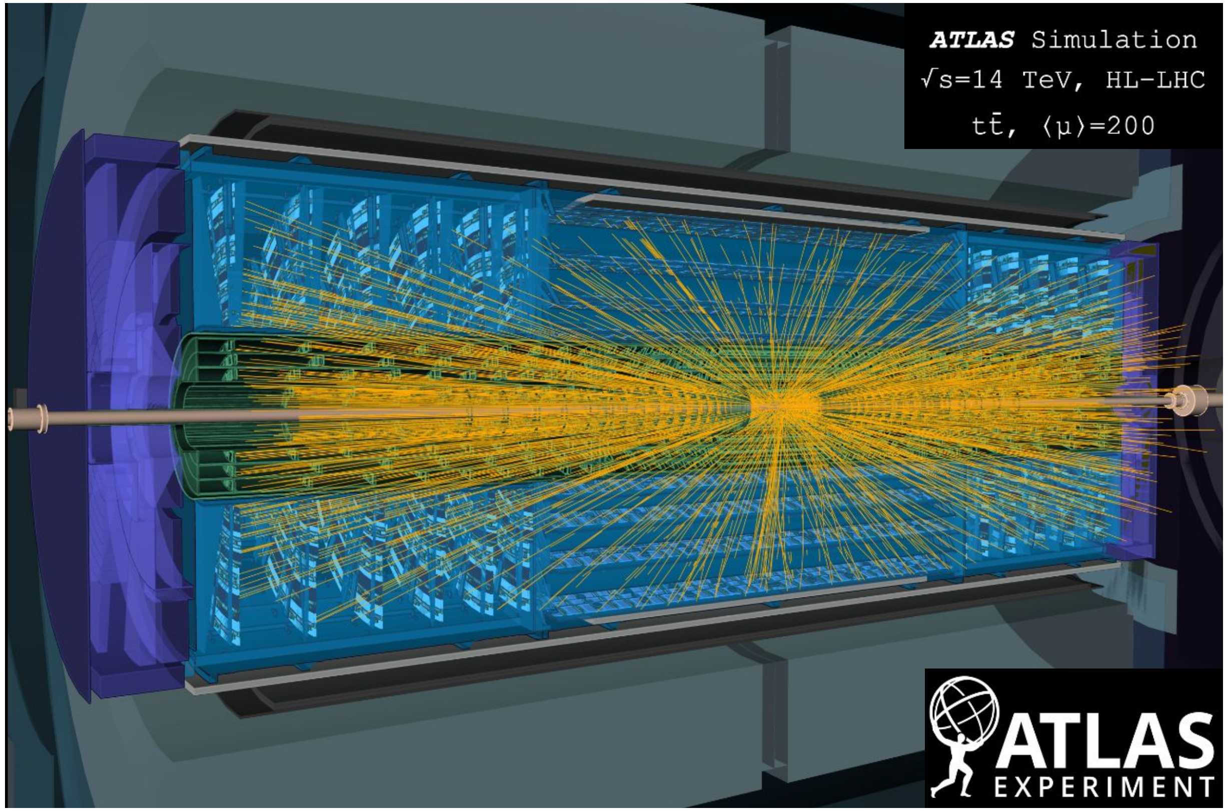
The ATLAS TDAQ Phase-II architecture



- **High-pileup conditions during the HL-LHC present a significant challenge to the ATLAS TDAQ system**
  - Extensive Phase-II upgrade program will encompass multiple systems, including tracking, triggering, and data acquisition.
- **Hardware trigger at Level-0: process data from subdetectors, identifies physics objects, evaluate event-level physics quantities**
- **Accepted events are transmitted to the Event Filter (EF):**
  - Provides the **final accept/reject decision**. If accepted, events are transferred for permanent storage.
  - Composed of **multi-threaded asynchronous processing farm** of **commodity servers** running a subset of **offline-like reconstruction algorithms** and **menu-driven event selection**
- EF farm will run **tracking**, **calorimeter** and **muon** algorithms

ATLAS Event Filter Tracking

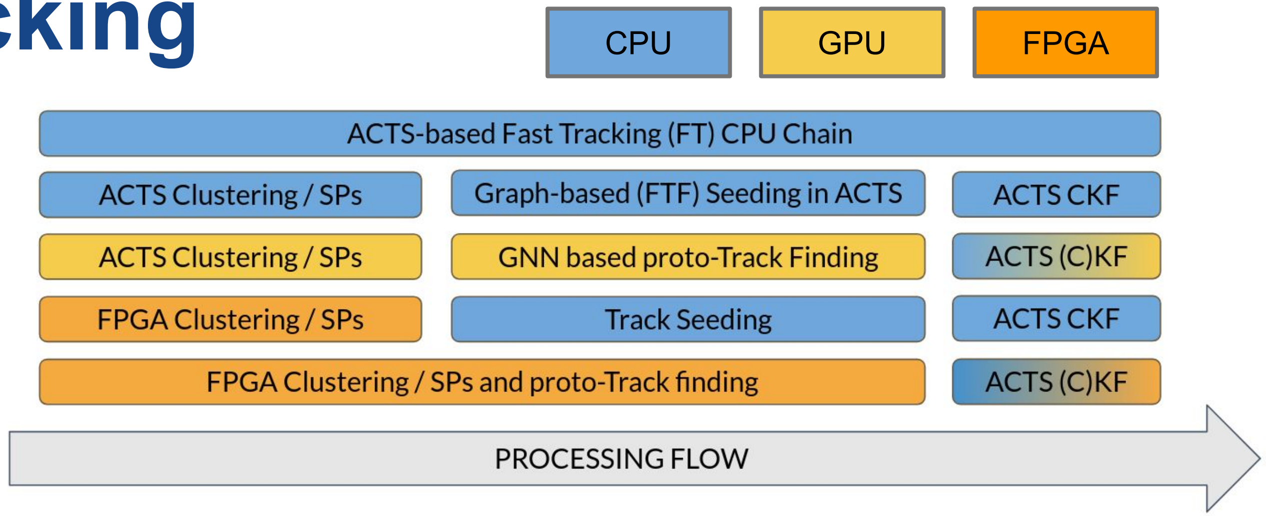
- **EF Tracking** refers to the **tracking algorithms running on flexible, heterogeneous commercial** system
  - TDR for Phase-II Upgrade of the ATLAS TDAQ System - Event Filter Tracking Amendment



- Consists of **ITk data preparation, track seeding and pattern finding, track fitting and ambiguity solving**
- Combines **Regional tracking** on the majority of the events, with **full-scan tracking** at a reduced rate
- Investigates **unconventional tracking strategies** for **new triggers**

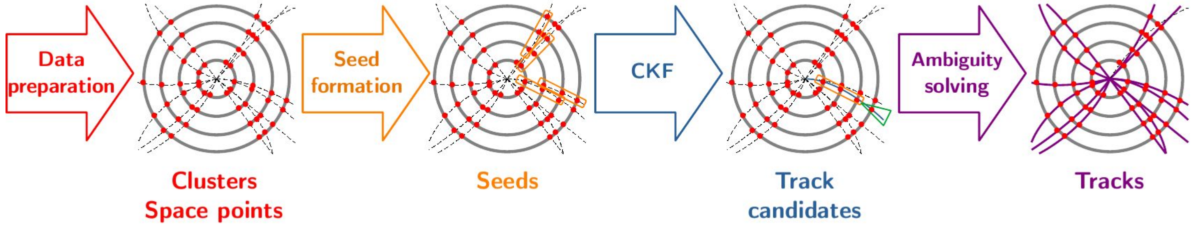
- **Biggest ongoing effort is deciding on the best technology for EF Tracking ➪ late 2025**
  - Exploring usage of CPUs, GPUs and FPGAs
  - Improving the algorithms, exploring new techniques

Task 2.4: Event Filter Tracking



- **Development of an algorithmic solution for the ATLAS EF track reconstruction**
  - Exploring **optimal classical numerical** and **Machine Learning** (ML) **techniques**, e.g. Graph Neural Networks (GNN)
  - Deployment on the **most suitable hardware architecture**: FPGA, CPU, GPU
- **Contributing to R&D lines** for algorithms and technologies to extend trigger event selection
- **Provide support and enable deployment** of the tracking chain at EF trigger
  - **Inputs for the 2025 technology choice report**

⇨ **Optimization of physics and processing performance of the track reconstruction at the EF**



- **Tightly connected to**
  - **WP 1.2**: Development framework towards fast inference of complex network architectures on LHC online systems
  - **WP 1.7**: Framework integration of accelerators
  - **WP 2.6**: Common Tracking Event Filter infrastructure

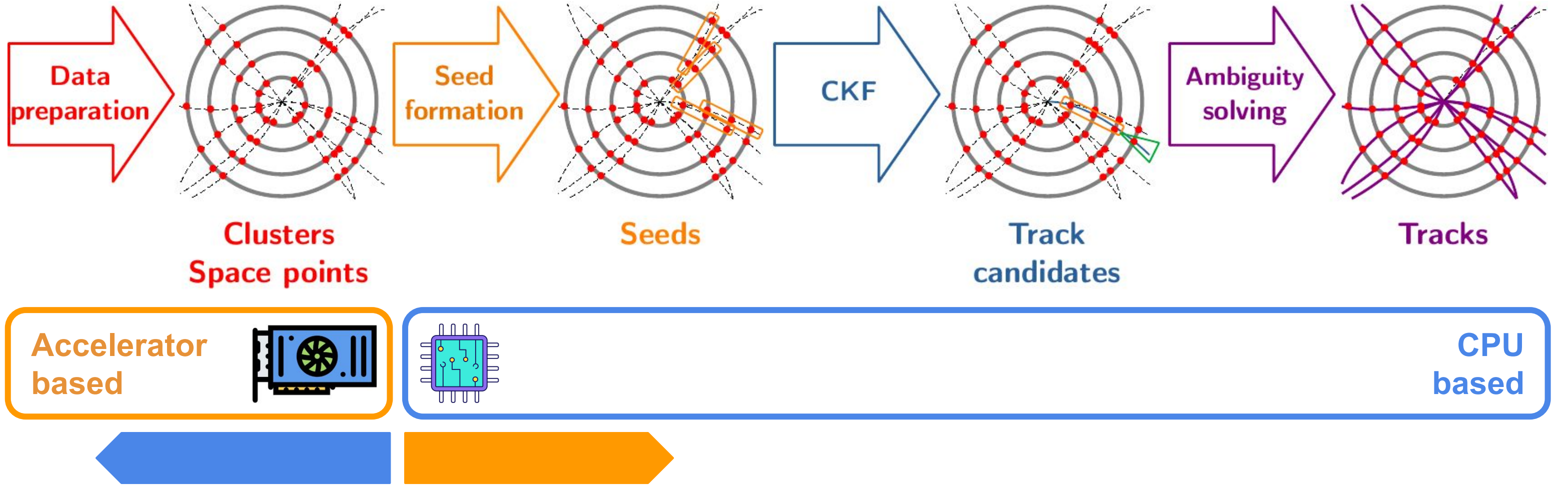
The team so far

- Task leaders: Noemi Calace (CERN), Stephanie Majewski (Oregon)
- Hiring started at the **beginning of the 2024** for positions **starting in June/July 2024**
  - LD Staff: **Pierfrancesco Butti**
  - 2 Applied Fellows (QUESTs): **Benjamin Huth** and **Julian Wollrath**
- **Selected students** from the last round of **Doctoral Student Program**
  - 2 Doctoral students will join in February 2025
- Personpower is by now very well **integrated into the ATLAS EF Tracking Project**
  - ATLAS authorship qualification projects and/or taking responsibility roles within EF tracking
  - … and not only EF Tracking: ID Tracking CP, ACTS core developers, GNN4ITk

Current activities in FPGA EF Tracking

- **FPGA implementation of each track reconstruction step**
  - **Multiple kernels** are developed and **deployed**
  - Certain steps are more efficient on CPUs than FPGA and vice versa
    - E.g. **full precision track fit** is almost **impossible to port to FPGA** due to the large memory requirement; **cluster formation** on FPGA can be **very performant**
  - Keep an eye on data transfer between CPU and FPGA
  - Contribution at CHEP 2024

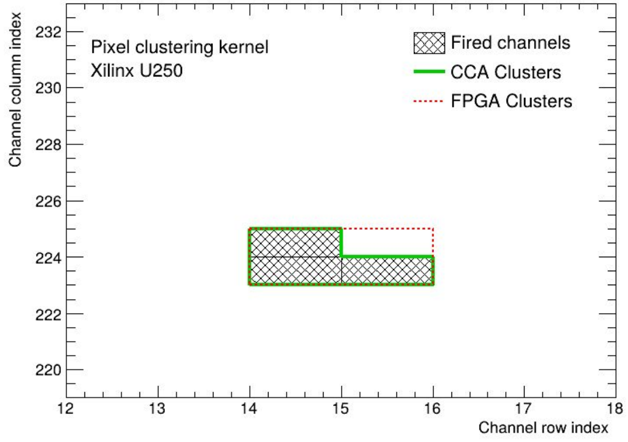
- Need to find where **boundary between FPGA and CPU** is for **best physics and computational performance**
  - Major contribution on ITk data preparation on FPGA

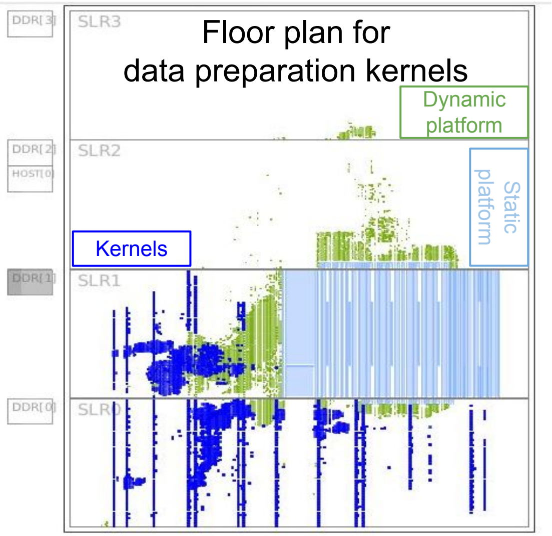


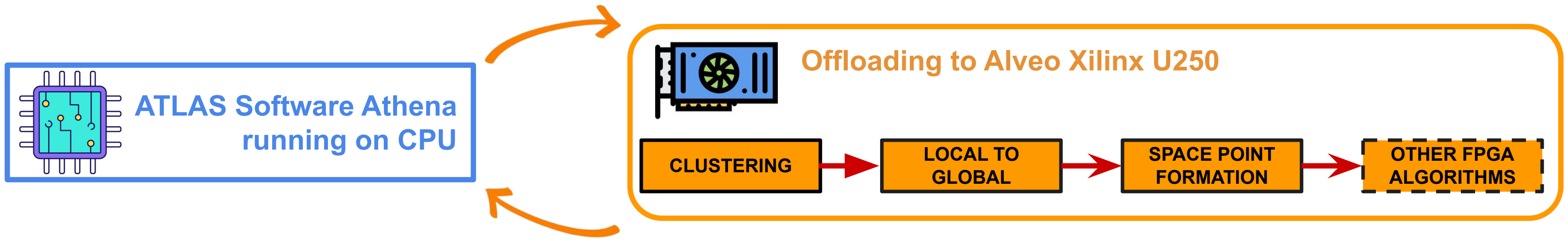
Sliding window which can be optimized for FPGA solution

Current activities in FPGA EF Tracking

- **ITk data preparation:** data from the detector is processed to create clusters which are used for tracking
  - Multi-step process with simple algorithms that can take advantage of FPGAs
  - Each step is implemented as separate a FPGA kernel within the Xilinx Vitis workflow
- ITk data preparation **FPGA kernels deployed and interfaced** with the **ATLAS software**
  - Established data-format for communication between kernels and between kernels and Athena
  - Testing and validation of the full set kernels for data preparation is ongoing
    - So far performance individually tested through test vector created from full physics simulation



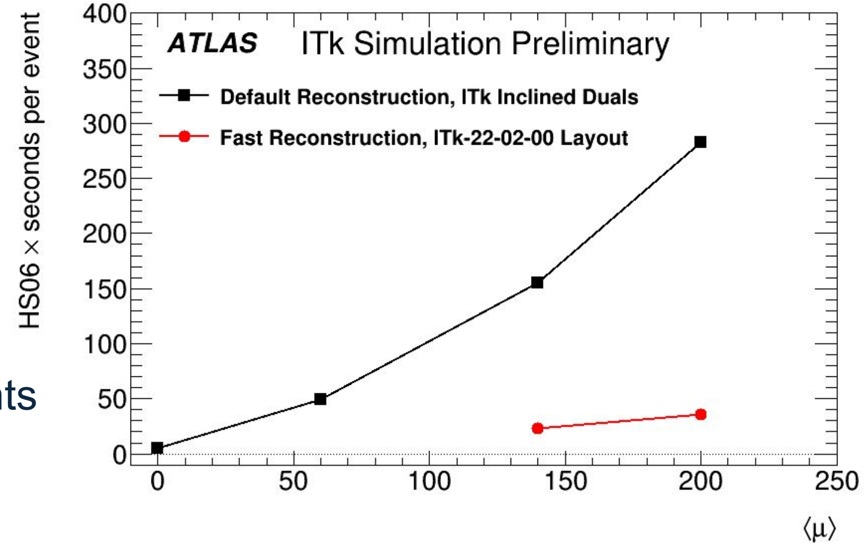




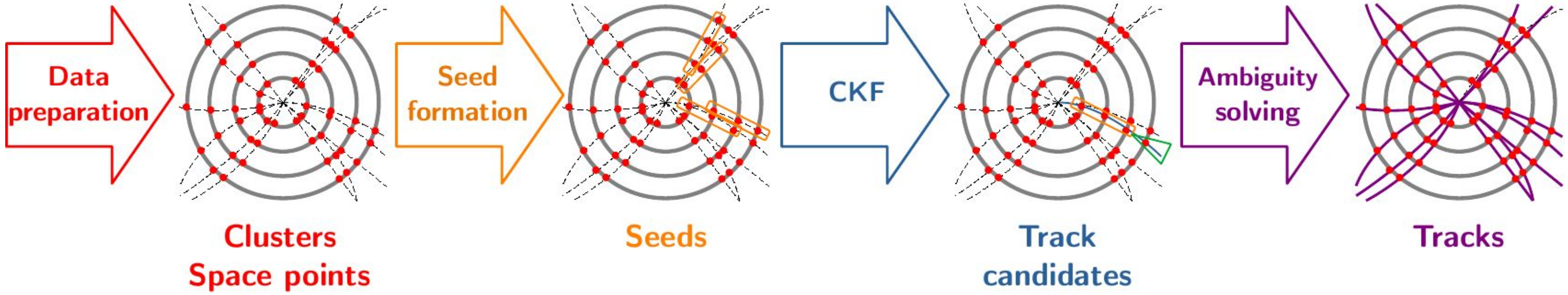
Current activities in CPU EF Tracking

- A **fully-functional** ITk **CPU-based *fast* track reconstruction prototype** for **EF Tracking** has been developed
  - **Baseline for the CPU-based solution** in ATLAS Phase-II TDAQ TDR Amendment
- **ATLAS** plans to make **extensive use of ACTS for Run 4**
  - Experiment-independent toolkit for track reconstruction in HEP experiments
    - Implemented in modern C++; thread-safe for parallelization/vectorization
    - Fully agnostic to detection technologies, detector design, and the event processing framework
    - Highly customizable and extendable ⇒ R&D platform

More from task 2.6

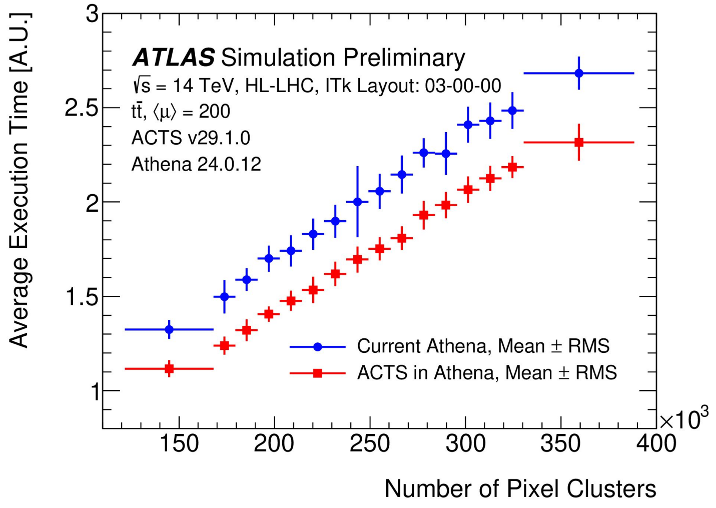


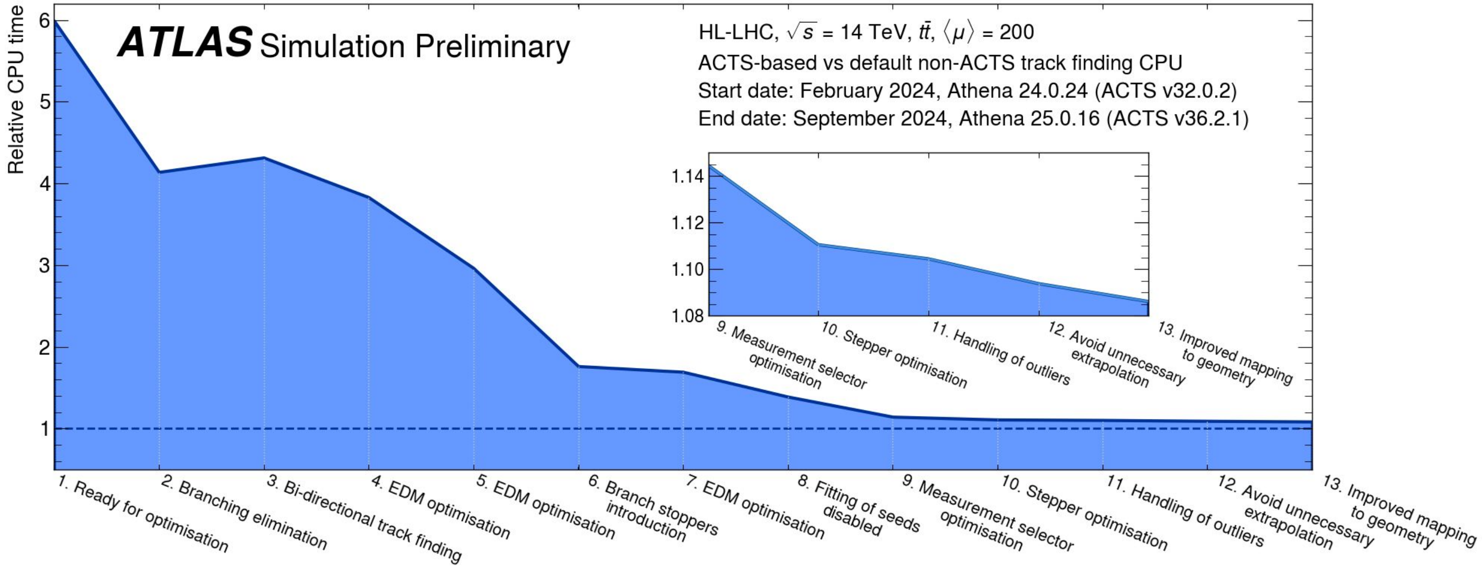
- Conceptually-identical **ITk Tracking chain** using **ACTS** toolkit and deployed **in ATLAS** reconstruction software
  - Identical or similar physics performance with **improvement on the computation side**!



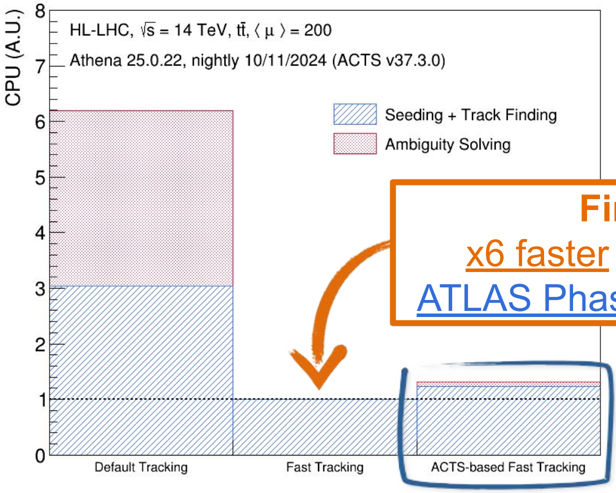
Current activities in CPU EF Tracking

- All components of the ITk tracking chain are **deployed and validated against athena counterparts [IDTR-2023]**
  - e.g. for clustering 10-20% speed up is achieved by ACTS implementation at ⟨μ⟩ = 200
- **Track finding** is based on the **ACTS Combinatorial Kalman Filter (CKF)**
  - Shows **very promising physics performance** results, and is **undergoing optimisation**
- **Very intensive ongoing optimization campaign of both ACTS and the ATLAS software**
  - Part of the ACTS integration effort
    - Contribution at CHEP 2024

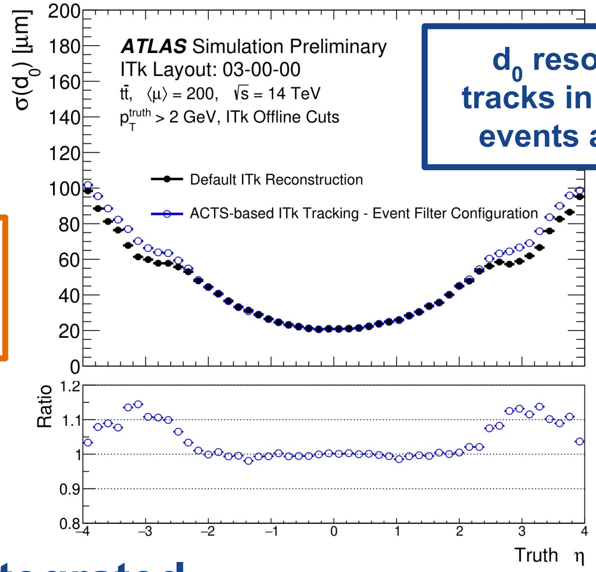




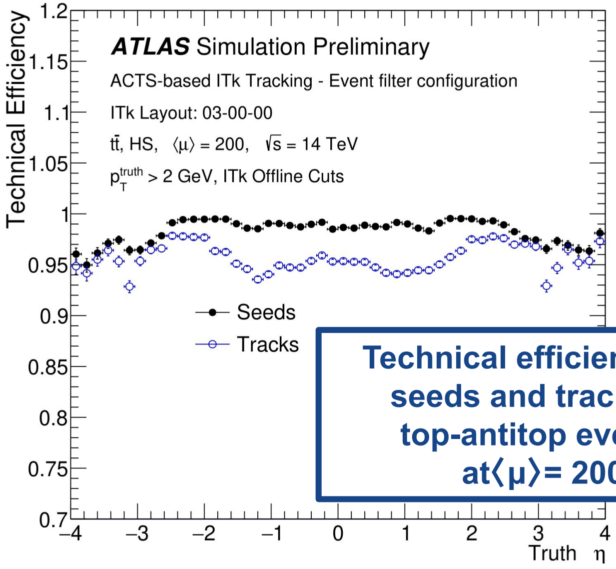
Current activities in CPU EF Tracking



First prototype
x6 faster than default tracking
ATLAS Phase-II TDAQ Amendment



d_0 resolution for tracks in top-antitop events at⟨μ⟩= 200



Technical efficiency for seeds and tracks in top-antitop events at⟨μ⟩= 200

- **CPU ACTS-based *fast* Tracking chain fully integrated**
  - x1.25 slower than the first prototype including ambiguity solving as well!
  - Very promising results already on tracking performance
  - Automatised monitoring for physics and computational performance
  - Execution using trigger chains and ROIs

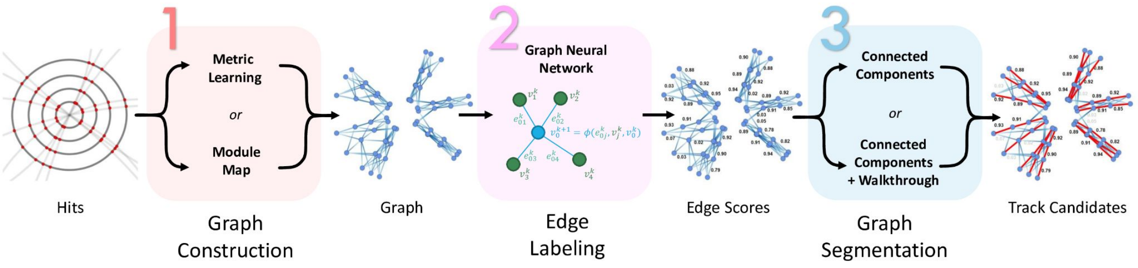
⇒ Keep improving physics and CPU performance

  - E.g. investigate alternative seeding possibilities

- **Very strong collaboration with task 2.6 on ACTS CPU optimization**

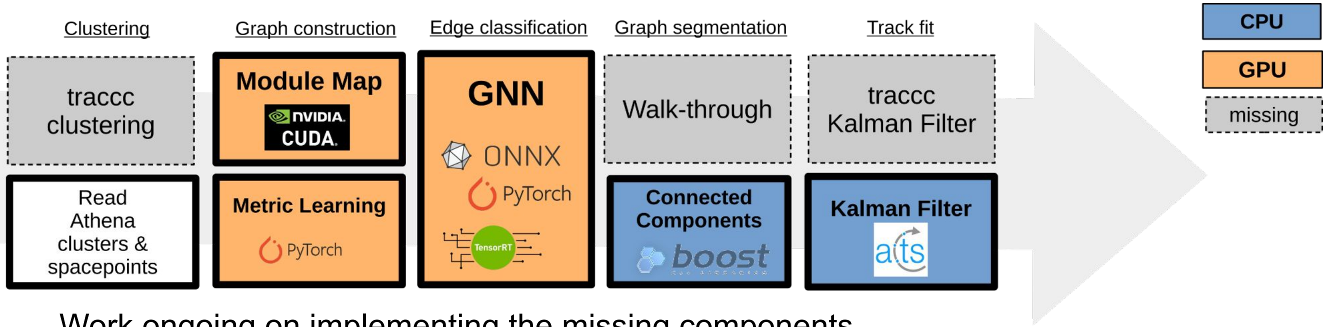
GNN-based solution on GPU for EF Tracking

- **ATLAS R&D line** on **tracking with GNNs** led by **ATLAS GNN4ITk group**
  - ATL-ITK-PROC-2022-006, IDTR-2023-06, several contributions to CHEP: talk, talk, talk, …



- **Strong involvement of NGT Task 2.4 in several areas**
  - **For the very first time in ATLAS**: investigation of **high performance inference frameworks** such as **Nvidia TensorRT** for GNN inference
  - Contributions to **CUDA-based graph construction library** of the offline GNN working group to facilitate **integration into tracking frameworks**

- Ongoing work on **designing the infrastructure** for performance studies
  - Standalone **ACTS-based GNN workflow** developed in collaboration **with task 2.6**
  - **ATLAS** simulation is used as **input data** to ensure realistic conditions
  - **Fuse with traccc** GPU-tracking code

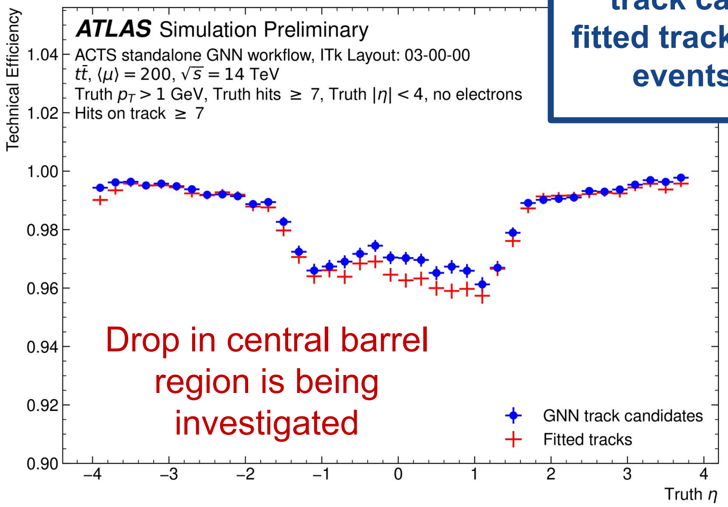


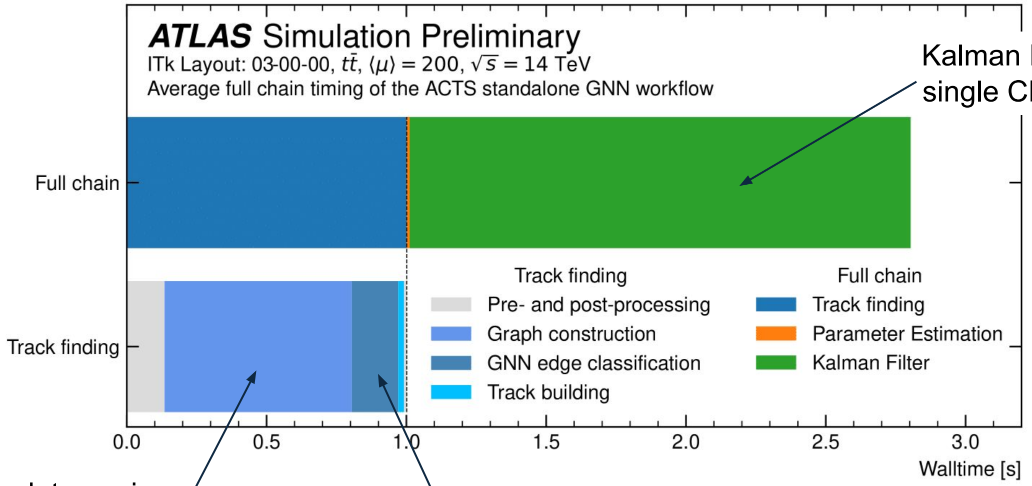
Work ongoing on implementing the missing components

GNN-based solution on GPU for EF Tracking

- **Preliminary tracking performance** is obtained using a **standalone ACTS-based GNN workflow**
  - Graphs are constructed using the “Module Map” approach
  - The GNN assigns a classification score *s* to each edge
  - Track candidates are built considering edges with score *s*>0.5 and using a connected-components algorithm

Technical efficiency for track candidates and fitted tracks in top-antitop events at⟨μ⟩= 200





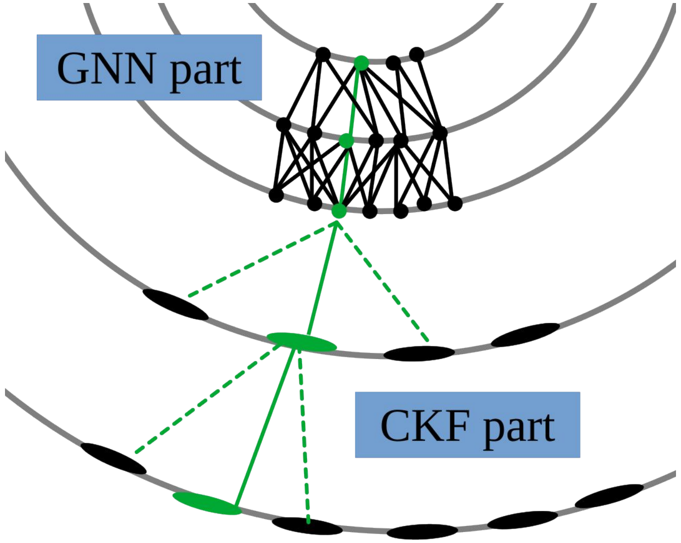
Kalman Filter on single CPU core

dominated by data copies, Core GPU algorithm takes O(100ms)

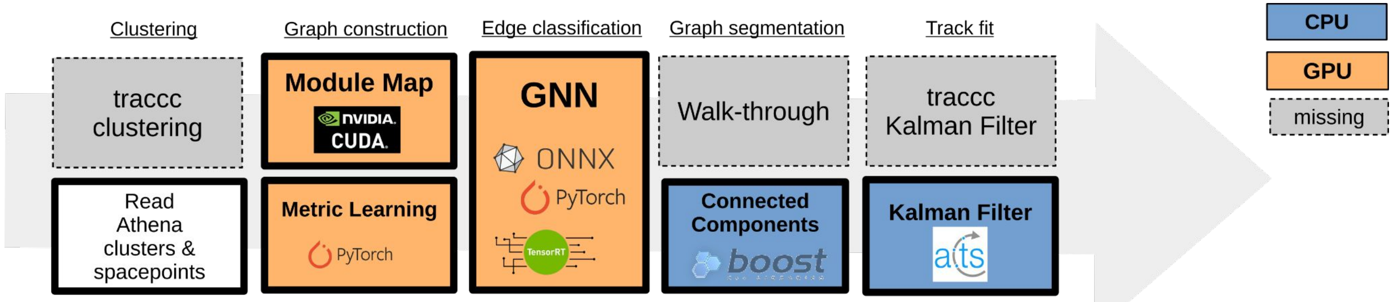
Nvidia TensorRT on GPU

- **Preliminary timing measurements with Nvidia A100**
  - **Promising timing** for individual steps
  - Chain not yet fully on-device (data copied to host after each substep)
  - GPU device not saturated, **no estimates of throughput yet**
    - Feedback and collaboration with WP 1.7

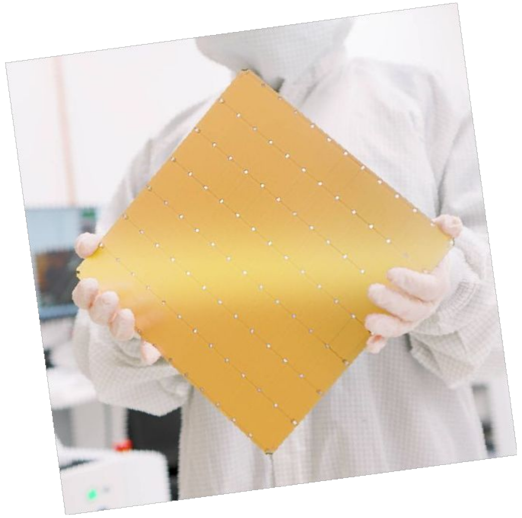
GNN-based solution for EF Tracking



- **More and more developments to come**
  - Enhance ACTS framework to support throughput measurements
    - In collaboration with task 2.6 and other work packages
    - Widespread discussion in ATLAS on how to compare different technologies
  - Integrate missing components (walkthrough, traccc fitter)
    - We will be able to use these functionalities directly in athena!
  - Use pruning & quantization to maximize GNN performance
    - Feedback will be provided to EF Tracking on using GNNs on FPGAs
  - Investigate **new strategies** and **new hardware**
    - Studies on pixel-only GNN tracking with CKF extension
    - Investigate new hardware (e.g. training on Cerebras AI chip)



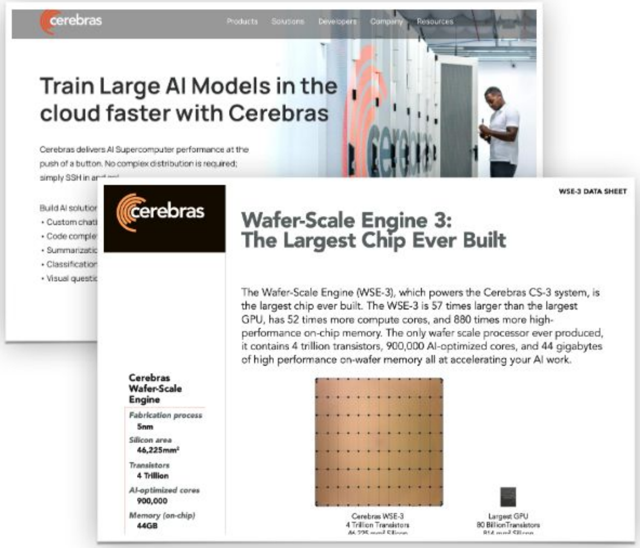
Work ongoing on implementing the missing components



Computing and Collaboration with WP1 and WP4

- **Active collaboration with WP1 and WP4**
  - Strong interest in results of WP1.7 for the support of heterogeneous architectures
  - Collaborating with WP1.2 and WP1.3 on HLS4ML and on the hardware aware ML architectures
    - Contribution to “hls4ml HEP Community Forum” as part of the *NextGen and SMARTHEP School on Edge Machine Learning*
  - WP1.1 is already providing us with hardware resources
    - ML training and GNN deployment studies: access to virtual machine with Nvidia A100 (40 GB) and AMD EPYC 7313 16-Core Processor, ~1.4 TB of disk space
  - We are already benefiting from the training opportunities provided by WP4 to our people
    - *NextGen and SMARTHEP School on Edge Machine Learning*, *Efficient Scientific Computing School 2024*

- **Work benefit directly from computing resources made available to us**
  - Beside the resources mentioned above, we also benefit from ATLAS and CERN TDAQ resources and test-beds at CERN
    - System administration for soon-to-be-installed EF-Tracking test-bench servers in TDAQ Lab4
  - Collaborating with Openlab to get access to further hardware architectures, e.g. Cerebras AI



Conclusion and next steps

- **2024 was a successful year with very important achievements!**
  - Hiring process, setting up activities, coordinating them with other work packages, define hardware needs, …
  - **Most important achievements**: NGT activities and personpower are integrated within the existing ATLAS structures especially **extending R&D** beyond baseline approaches!

- **Agenda plenty of interesting developments in front of us!**
  - NGT fuel is the **expertise of the entire community**: strongly benefits from collaboration within and across work packages!

CERN

NexTGen

Next Generation Triggers