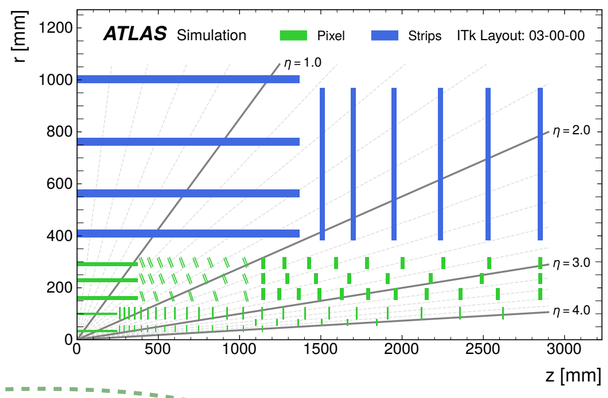
Regional reconstruction of tracks for the ATLAS Event Filter using GPU accelerators

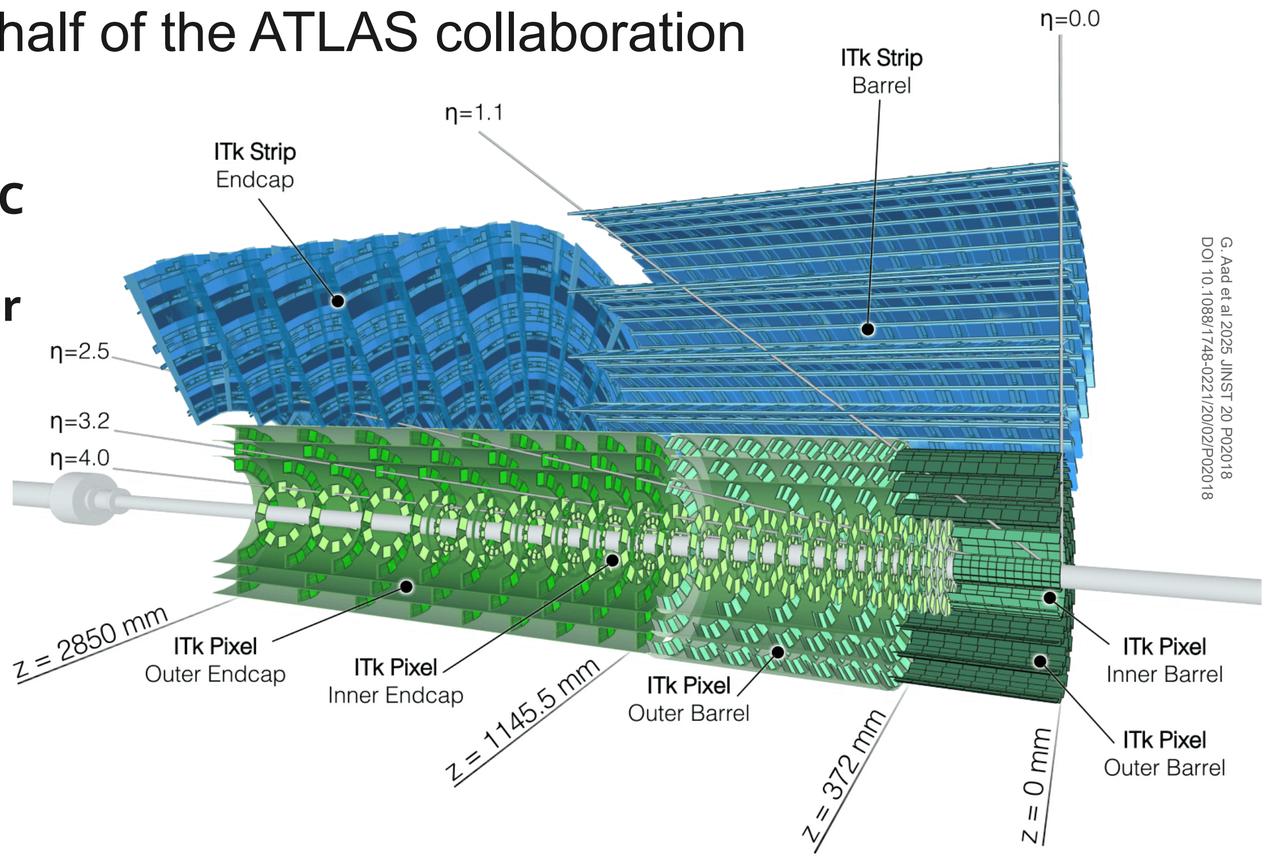
Benjamin Wynne, on behalf of the ATLAS collaboration

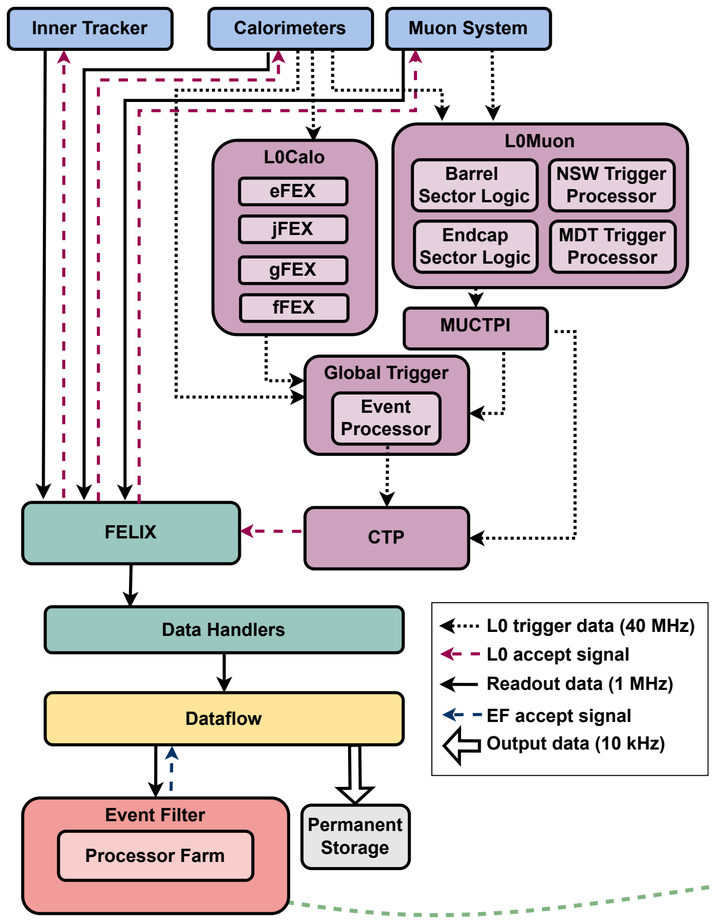
G. Aad et al 2025 JINST 20 P02018
DOI 10.1088/1748-0221/20/02/P02018

The upcoming high-luminosity upgrade to the LHC will increase the average number of concurrent proton-proton collisions <μ> in the ATLAS detector from ~60 to ~200.

An upgraded tracking detector (ITk) will be installed, and a new Event Filter (EF) processor farm will be commissioned with GPU accelerators.

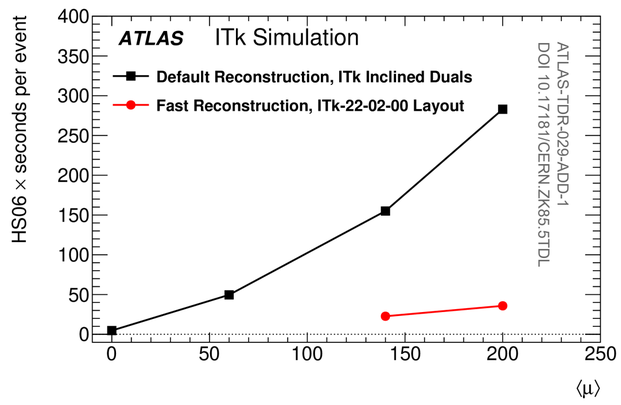






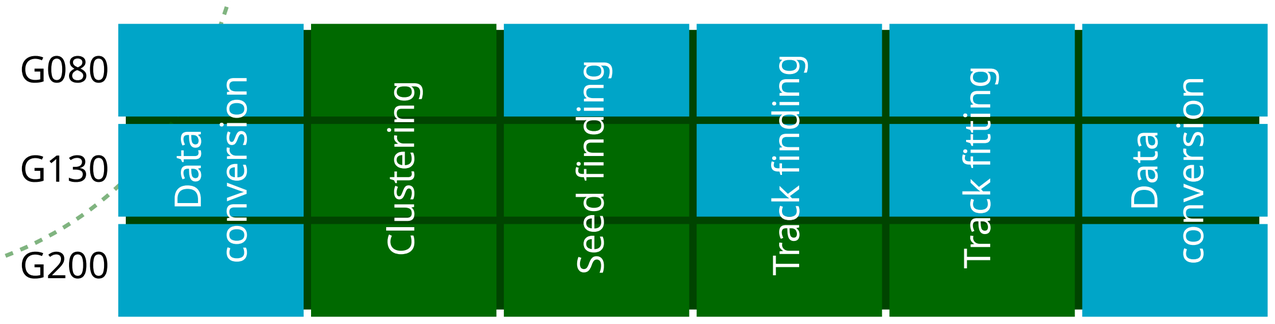
Track reconstruction is a combinatoric problem, with increased detector hits and track candidates leading to a non-linear scaling of compute requirements.

Algorithmic improvements have been demonstrated, particularly a graph-based approach to track seeding, but the intention is also to offload the most parallelisable tasks to GPU, with the rest performed by ACTS [1] CPU algorithms.



The GPU implementation is provided by the traccc [2] project, with portable kernels for each stage of the tracking pipeline allowing for multiple prototypes with different amounts of offloading. Full prototypes for CPU-only and CPU+FPGA processing were also created and evaluated.

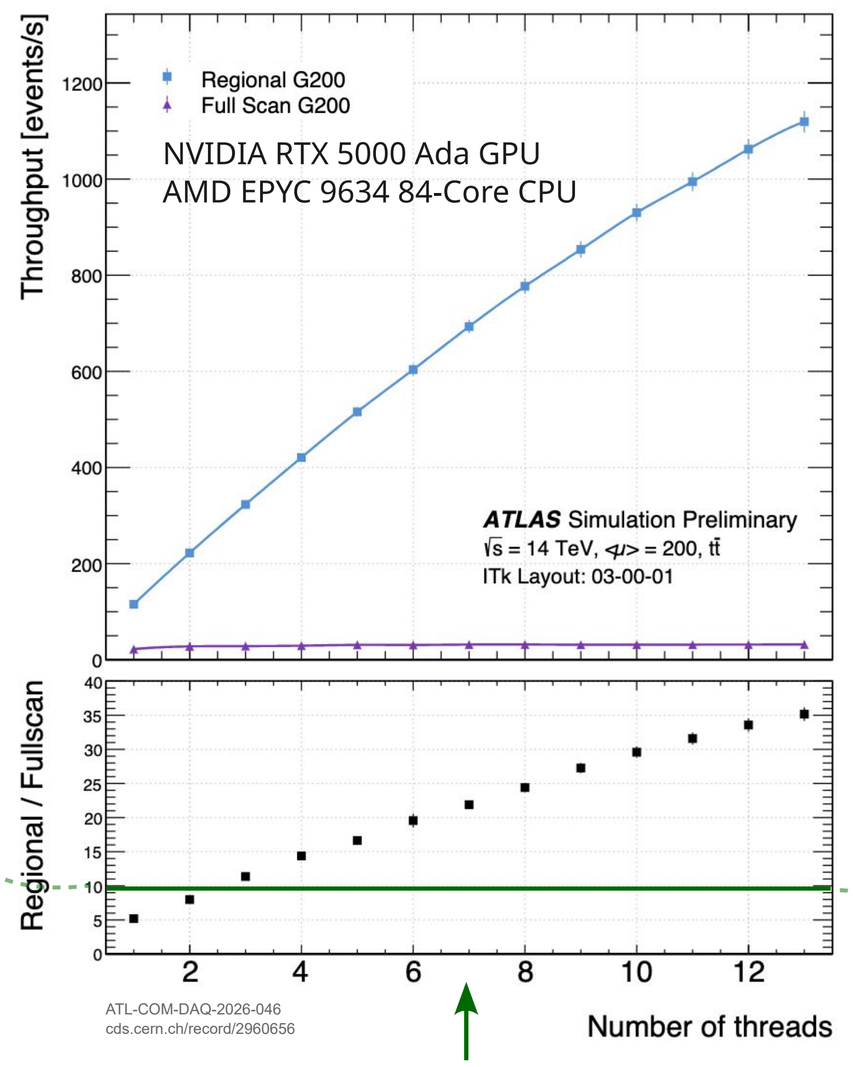
[1] arXiv:2106.13593
DOI 10.48550/arXiv.2106.13593
[2] arXiv:2505.22822 [hep-ex]
DOI 10.48550/arXiv.2505.22822

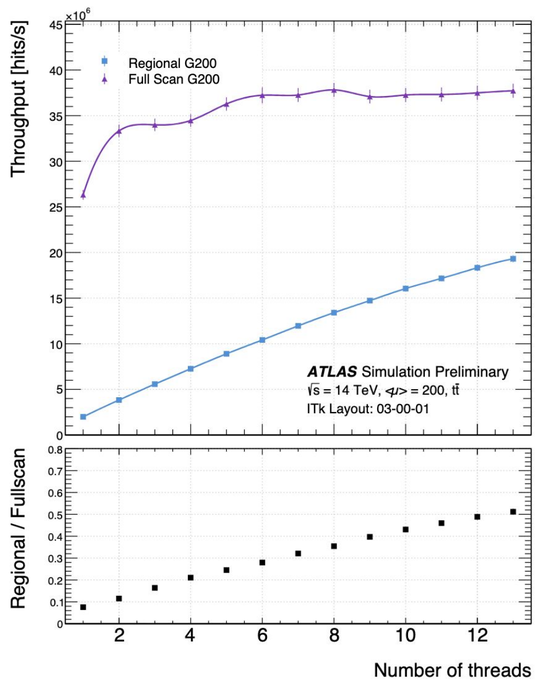


The EF is required to process events at 150 kHz, but also to reconstruct "Regions of Interest" (RoIs) comprising ~5% of the full detector space, at a rate of 1 MHz.

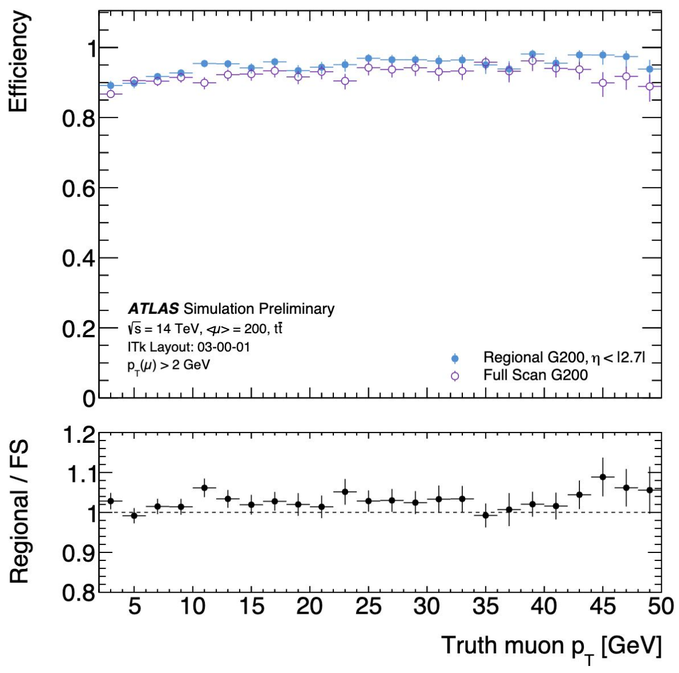
Cost projections use 100 kHz of event processing as a proxy for 1 MHz of RoIs. For this to be valid we must demonstrate that we can process RoIs 10 times faster than full events.

In this prototype, traccc loads a single copy of the full ITk geometry into GPU memory, and then processes many RoIs in parallel, where each RoI is provided as the set of hits from the corresponding detector region.





The scaling behaviour is explained by the data volume in terms of detector hits. For full events the GPU saturates, whereas for the much smaller RoIs that ceiling is not reached.



Track reconstruction efficiency in RoIs is consistent with full events (small improvement is due to bias from RoI creation).

Saturating this single GPU requires multiple CPU threads, with the specific CPU:GPU ratio depending on the relative load of full event processing versus RoIs, and also the amount of offloading (G200 represents maximal GPU workload). Final system architecture may require flexible provisioning of devices as-a-service.

Track reconstruction in Regions of Interest can be achieved at HL-LHC target rate with existing ATLAS GPU prototypes.