Demonstration of FPGA-based track reconstruction on live LHCb data

Why FPAGs for track reconstruction?

- LHCb reconstructs events at 30 MHz in Run 3 ($\mathcal{L} = 2 \times 10^{31} \text{cm}^{-2}\text{s}^{-1}$)
  - High Level Trigger: Level 1 (HLT1) on CPUs + Level 2 (HLT2) on CPUS
  - Expected factor $x5$-10 in luminosity with Upgrade II\(^1\) presents a challenge
- HEP experiments are seeking heterogeneous computing solutions in view of increasing luminosity with Moore’s law slowing down\(^2\)
- Modern FPAGs can perform highly parallel processing with high throughputs and low latencies
- FPAGs as greener solution: less power-hungry than CPUs and GPUS

The “Artificial Retina” architecture

- “Artificial Retina” architecture\(^3\) is a fast implementation of a computation resembling the Hough transform\(^4\) and deployable on FPAGs\(^5\)
- Generalised approach (not only straight lines as track model)
- Numerical evaluation: using a preset of reference tracks
- Continuous (non binary) response quantifying agreement with reference tracks and subsequent interpolation

The system implementation

- Each cell of the matrix implements an engine for computing the weights
- Engines work in fully parallel way, thus FPGA-friendly
- System can be spread over multiple FPAG boards
- Distribution network for spreading data across the system
- Switch for distributing hits to cells where their weight is significant
- Fast optical network for inter-board communication\(^6\)

Demonstrator

- Installed at the Coprocessor TestBed facility located at LHCb site
- 8 Intel Stratix 10 FPAG
- LHCb VELO quadrant coverage
- Full-mesh network for fast data exchange
- 28 full-duplex links at 25.8 Gbps
- Engines on different boards cover different parameter space regions

Additional tests and future prospects

- Official LHCb Montecarlo simulated events, with Run 3 conditions ($\mathcal{L} = 2 \times 10^{31} \text{cm}^{-2}\text{s}^{-1}$), are injected in the demonstrator internal RAMs and read in loop in order to maximise input rate
- Comparison between tracks reconstructed by the hardware and custom deployed C++ emulator shows exact bitwise adherence between the two
- About 10 days of running without errors (much higher than bunch life)

The event throughput has now reached an unprecedented rate of 19 MHz

A final rate of ~31 MHz is estimated when all ongoing tunings will be completed

Proposal for a FPGA-based downstream tracking in Run 4 is currently under review by the LHCb collaboration

Processing live LHCb data - Data flow

- Deployed custom chain for feeding data from LHCb DAQ in real-time to the demonstrator and apply most recent alignment at every new run
- Communication with FPAGs via PCIe

Processing live LHCb data - Results

- Run more than 30 days of pp-collisions data without hiccups with data delivered to the TestBed facility at the event rate of 1 kHz
- Real-time reconstructed tracks appear sensible when their distribution is qualitatively compared with LHCb HLT2 reconstruction
- Retina Architecture demonstrator is able to reconstruct a portion of a detector in real time with real data coming from the detector DAQ

These results are achieved thanks to the funding from Italian INFN and the kind support of the RTA and Online groups in LHCb

\(^1\) LHC-B Collaboration, “Expression of Interest for a Phase-II LHCb Upgrade: Opportunities in flavour physics, and beyond, in the RIC-VEC era”, COMPENDIA-2012-001 (2012)
\(^3\) L. Ristori, “The artificial retina for fast track finding”, Doctorate Thesis, 2017
\(^5\) G. Ton and P. Frix, “Reconstruction of track candidates at the HLT using FPGAs”, LHC White Card 946, 0300 (2016)
\(^6\) F. Lazzari et al., “FPGA-based real-time data processing for outstanding reconstruction of LHCb”, Journal of Instrumentation 17, 04003 (2022)