Nuclear Instruments and Methods in Physics Research A 1069 (2024) 169782

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, A

journal homepage: www.elsevier.com/locate/nima

Detector-embedded reconstruction of complex primitives using FPGAs

Giovanni Punzi [b,i,*], Wander Baldini [a], Giovanni Bassi [b,c], Andrea Contu [d], Riccardo Fantechi [b], Jibo He [e,f], Brij Kishor Jashal [g], Sofia Kotriakhova [a,h], Federico Lazzari [b,i], Maurizio Martinelli [j,k], Diego Mendoza [g], Michael J. Morello [b,c], Arantza De Oyanguren Campos [g], Lorenzo Pica [b,c], Qi Shi [e], Francesco Terzuoli [b,l], Giulia Tuci [m], Ao Xu [b], Jiahui Zhuo [g]

[a] INFN Sezione di Ferrara, Ferrara, Italy
[b] INFN Sezione di Pisa, Pisa, Italy
[c] Scuola Normale Superiore, Pisa, Italy
[d] INFN Sezione di Cagliari, Monserrato, Italy
[e] University of Chinese Academy of Sciences, Beijing, China
[f] Hangzhou Institute for Advanced Study, UCAS, Hangzhou, China
[g] Instituto de Fisica Corpuscular, Centro Mixto Universidad de Valencia CSIC, Hangzhou, Spain
[h] Universita' di Ferrara, Ferrara, Italy
[i] Universita' di Pisa, Pisa, Italy
[j] INFN Sezione di Milano-Bicocca, Milano, Italy
[k] Universita' di Milano Bicocca, Milano, Italy
[l] Universita' di Siena, Siena, Italy
[m] Physikalisches Institut, Ruprecht-Karls-Universitat Heidelberg, Heidelberg, Germany

ARTICLE INFO

Keywords:
DAQ
FPGA
Trigger
LHCb

ABSTRACT

The slowdown of Moore's law and the growing requirements of future HEP experiments with ever-increasing data rates pose important computational challenges for data reconstruction and trigger systems, encouraging the exploration of new computing methodologies. In this work we discuss a FPGA-based tracking system, relying on a massively parallel pattern recognition approach, inspired by the processing of visual images by the natural brain ("retina architecture"). This method allows a large efficiency of utilisation of the hardware, low power consumption and very low latencies. Based on this approach, a device has been designed within the LHCb Upgrade-II project, with the goal of performing track reconstruction in the forward acceptance region in real-time during the upcoming Run 4 of the LHC. This innovative device will perform track reconstruction before the event-building, in a short enough time to provide pre-reconstructed tracks ("primitives") transparently to the processor farm, as if they had been generated directly by the detector. This allows significant savings in higher-level computing resources, enabling handling higher luminosities than otherwise possible. The feasibility of the project is backed up by the results of tests performed on a realistic hardware prototype, that has been opportunistically processing actual LHCb data in parallel with the regular DAQ in the LHC Run 3.

1. Introduction

The historical evolution of experiments has always been going hand-in-hand with the increase of data processing rates. However, some experiments have larger demands than others, due to the physics they address. Amongst them LHCb is a notable example, addressing flavour physics at low-Pt and very large rates, leading it to need to process the largest data flow in the field, both currently and in the foreseeable future, topping at around 200 Tb/s in the planned Upgrade 2 for the LHC Run 5. This need drives LHCb to keep searching for new processing solutions. The current Data Acquisition system of LHCb is schematically represented in Fig. 1. The scheme is based on a triggerless readout of the whole detector at the LHC crossing rate, to perform full event reconstruction before making the first trigger decision (HLT1, GPU-based). The final event selection occurs in the next stage (HLT2), based on a second and final reconstruction, that is persisted offline for all physics purposes. Detector calibration and alignment occurs between

* Corresponding author at: Universita' di Pisa, Pisa, Italy.
E-mail address: giovanni.punzi@cern.ch (G. Punzi).

https://doi.org/10.1016/j.nima.2024.169782
Received 30 June 2024; Received in revised form 20 August 2024; Accepted 21 August 2024
Available online 28 August 2024

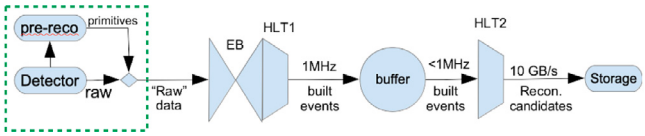

Fig. 1. LHCb DAQ diagram. The box highlights the proposed addition.

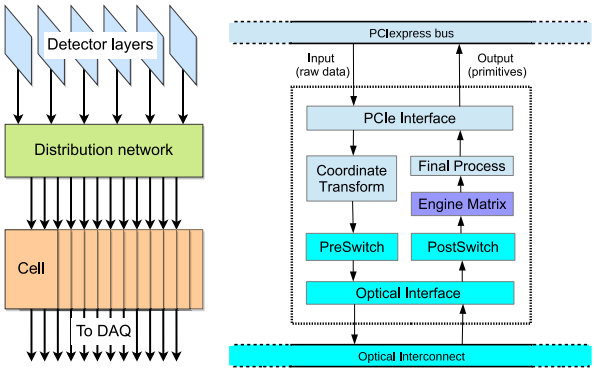

Fig. 2. Data flow in the Retina architecture, and its mapping to boards.

HLT stages, taking advantage of a large disk buffer. While this scheme works well for the current Run 3, the further order of magnitude rate increase planned for Run 5 calls for a further evolution.

In the present work, we describe a proposed enhancement to this scheme, based on the pre-reconstruction of intermediate data structures ("primitives") to accelerate reconstruction in the following stages and reduce data flow. The underlying idea is to perform this step at a very early level, before full events are built, using local detector information. The goal is to make this transparent to the rest of the DAQ, embedding it in the readout (Fig. 1) and formatting primitives as raw detector data. While this strategy has already been used in limited ways in both LHCb [1] and other experiments, the proposal is now to make this a major part of the event reconstruction process, and persist this up to the offline analysis level. The most appropriate technology for this evolution appears to be the use of current FPGA devices, that provide flexible processing power with low latency, low power, and large bandwidths — all at affordable prices.

2. Principles of operation

The technology at the core of our system is a highly-parallel architecture for pattern recognition that was originally named "artificial retina". It owes its name to the effort of mimicking some structural features and general principles found in the organisation of the natural neural network responsible for fast vision processing in living organisms [2]. The *Retina* architecture was designed for real-time pattern recognition with very low latencies, avoiding any buffering or time-multiplexing, just as in natural vision systems. This is obtained by a design focused on extreme parallelism and generous use of bandwidth resources [3–5]. These features match well the requirements of the reconstruction scheme described in the previous section.

The Retina architecture is an arrangement of parallel computing units fed by a custom switching network (Fig. 2), programmed to perform a computation resembling the "Hough transform" [6]. Its purpose is to find patterns in its input, arriving on multiple parallel lines, that are compatible with a set of pre-calculated reference patterns. The output is similarly provided on multiple parallel lines, to avoid any bottleneck due to serialisation (Fig. 2). The throughput of this architecture can be very high when implemented with low-latency components and a large bandwidth switching network, as provided by modern FPGA devices. However, even the largest FPGAs on the

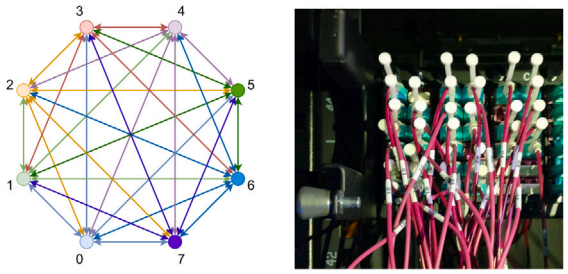

Fig. 3. Connection map, and zoom of patch-panel cabling for the 8-nodes full-mesh network.

market are not large enough to handle an entire LHCb subdetector. This has been solved by a vertical segmentation, both of the distribution network and the cell matrix, breaking down the system into smaller blocks that can be fitted into individual FPGAs, interconnected by fast optical point-to-point links, and performing I/O through PCIe interface to a host server (Fig. 2).

In functional terms, the chosen parameter space is discretised, and represented in a matrix of cells. Each cell is mapped to a custom computational engine implemented in a modest amount of logic circuitry inside the FPGA, physically independent from all others. The task of the engine is to receive hits in input and check their compatibility with the assigned reference track. To this purpose, each computational engine assigns a weight to every received hit, that is determined by its distance from the reference track on that layer. The weights of all received hits are accumulated into a single value, whose final value ad the end of the event represents the 'excitation level' of the corresponding cell. A high value indicates good compatibility of the set of received hits to the reference track.

After all cells have received all relevant hits from the event at hand, the next step of cluster finding begins. This is the determination of local maxima of the excitation function within the matrix of all cells; they are taken as track candidates, with the position of their centroid within the matrix being an estimate of the candidate track parameters.

3. Hardware demonstration

The practical viability of the complex system described in the previous section can only be demonstrated by a real hardware implementation. We opted to build a life-size demonstrator, covering a significant portion (~1/4) of the LHCb VELO detector, with the complete functionality expected from the final device. The demonstrator is made of 8 PCIe Gen3 boards, each carrying a Stratix 10 FPGA, produced by Bittware/Molex and commercialised under the name 520N. They have a PCIe Gen3 x16 connector and 16 XCVRs with a maximum bandwidth of 26 Gb/s each. The demonstrator covers 16 VELO modules of the A-side (even-numbered modules from 16 to 46), within a quadrant of track parameters space and is contained within a single server of the same make and model of the Event Builder servers, to ensure full compatibility with the LHCb environment.

The inter-board communication occurs over a 8-node full-mesh network, that was implemented with 3 breakout cassette and 8 fan-out cables, resulting in a very compact Patch Panel (Fig. 3). Extrapolation of this scheme to the full VELO detector, considering all the necessary extra connections, implies a Patch Panel with 22 breakout cassette and 64 fan-out cables, arranged in a 'dragonfly' topology, that can be fitted in just 4U of rack space. The achievement of a compact design is a crucial feasibility proof for future larger systems.

For serial inter-board communication, we adopted the Intel Super-Lite II V4 protocol, an open-source, highly efficient protocol (96.3%) that provides a flow-control mechanism for implementing back-pressure, and allows to connect each XCVR to a different endpoint, thus fulfilling all requirements of the Retina architecture.

The demonstrator is installed at the LHCb Coprocessor TestBed facility (Fig. 4) to emulate realistic data taking conditions, and has been subjected to several types of tests.

During LHCb physics data taking, the demonstrator is made to process real data produced at LHCb collision point, in parallel with the regular DAQ and in real time. To achieve this, raw events from the LHCb monitoring farm are buffered and stored on disk, in files of approximately 2 GB each. VELO RawBanks are extracted from the file and decoded to retrieve the hits (produced by the cluster-finder firmware [1]), that are then fed via PCIe to the demonstrator boards, each of them covering two VELO modules. The hit loading is carried out without any time constraints on event synchronisation; nonetheless, continuous internal monitoring of the system showed that no corruption or mixing of events has ever occurred during months of testing. The output FIFOs of all boards are read and the reconstructed tracks stored onto disk, preserving Run Number and the original name of the incoming file to allow offline analysis. While only a subsample of the events can be transmitted in this way (about 1 kHz) due to intrinsic limits in data transfer to the testbed facility, this setting allows to test the behaviour of the system when faced with the variable and unpredictable conditions of real data taking. This includes changing alignment conditions, that are accounted for by applying to the VELO hit coordinates the most recent alignment constants, after retrieving from the same database used by the main LHCb DAQ.

This test was run extensively during data-taking with *pp* collisions in mid July and September 2023. The demonstrator ran smoothly for the entire period, and the outputs showed no errors or any other anomalous behaviour. The analysis of the output data show a distribution and rate of VELO tracks that is consistent with what is found by the standard reconstruction, and tracked reliably also with large changes in VELO positioning. The lack of persistency of the event numbers, and the limited rate, prevent performing more quantitative tests in this modality, and motivated further tests that have been performed separately.

The bit-by-bit accuracy of the processing has been tested periodically during the real data run, by artificial injection of data packets inside the same real-data processing chain, taking advantage of the low input rate from LHCb, that leaves a lot of unused processing time. This procedure allows detailed monitoring of the functionality of the device over time, and has always returned results perfectly consistent, at the bit level, with the results of the detailed software emulation of the demonstrator, available as a C++ piece of code. This ensures the strict adherence of the demonstrator hardware to the intended processing.

Extensive tests have separately been performed on simulated minimum bias events from the LHCb official simulation, generated at an instantaneous luminosity of 2×10^{33} cm^{-2} s^{-1}. Data is loaded in circular buffers in the internal RAMs of the boards, allowing prolonged testing at maximum speed. While running on these simulated data, the output of the demonstrator has also been continuously checked against emulation predictions at the bit-level, but the large throughput in this configuration allows only a random subsample of the output to be tested.

The system was able to run without interruptions or errors for long periods (up to 27 days), and has only been stopped for external causes. This was repeated several times. Errors have been detected only in one occasion, when the temperature of the setup had been allowed to raise significantly above the nominal operating range of its components, specifically the serializers/transceivers.

During those extensive tests at top speed, we measured the total power consumption through the available on-board sensors. This measurement is performed on the board power lines and therefore includes the power consumption of the FPGAs, of the QSFP modules, and of the boards themselves (the patch-panel is passive and does not require power). The total power absorbed by the 8 demonstrator cards, when operated at their maximum rate, was measured to be 0.55 kW, which compares well with other systems under similar computing loads [7].

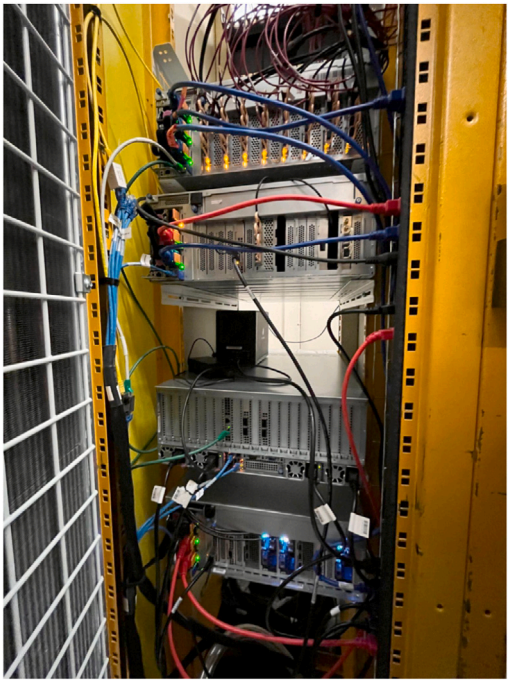

Fig. 4. Photo of the Retina demonstrator in the LHCb Coprocessor TestBed. The server holding the 8 boards is visible at the top, with optic cables running towards the patch panel located above.

However, the most important parameter measured in this test is arguably the event throughput, that was measured in the above tests to be 19.6 MHz. This is an unprecedented rate of event reconstruction by a single (non-multiplexed) device, for events as complex as those coming from the LHCb VELO detector, at the Run 3 nominal luminosity of 2×10^{33} cm^{-2} s^{-1}. While still falling a bit short (82%) of the ~24 MHz colliding bunches rate at LHCb in the LHC filling scheme of Run 3, it is a clear proof of the feasibility of performing tracking primitive reconstruction at level zero with this approach. In fact there are a number of tunable firmware parameter that can be optimised, leading to an estimated throughput increase of more than a factor of 1.6x achievable already with the current hardware, thus exceeding the crossing-rate requirements. In addition, the FPGAs used in our demonstrator are today already surpassed in performance by other devices currently on the market, that will be the minimal target of any future applications (e.g. the Stratix 10 was superseded by the *Agilex* series, as its direct successor in the same family).

Of greater importance for future application is instead the scaling of the throughput with increasing luminosity/occupancy. We performed specific tests of this behaviour, by merging n events simulated at Run 3 luminosity into a single one, to mimic the effects of luminosities a factor n higher. Tests have shown that the inverse throughput of our demonstrator scales linearly with the instantaneous luminosity, up to values higher than the maximum that LHCb is ever expected to reach, and conversely that the throughput increases linearly with the size of the hardware (number of cells). This confirms and puts on solid grounds the expectation based on the architectural design of the system, of a processing cost growing just linearly with event occupancy. This sets this approach apart from most other computing solutions, whose cost typically grows with larger powers of the occupancy due the combinatorial nature of the track-finding problem, and makes it a particularly promising solution for high intensity environments.

4. The downstream tracker project

Following the studies described in the previous sections, a project has been approved by the LHCb collaboration, to implement a device

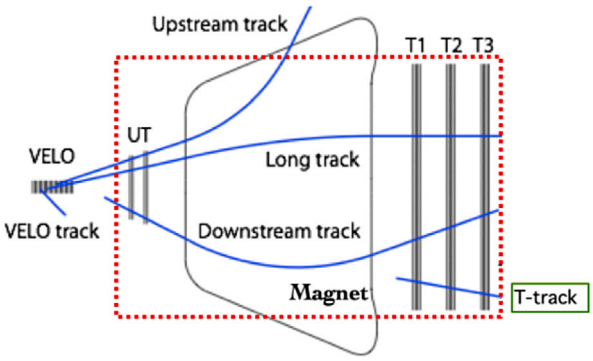

Fig. 5. Scheme of the LHCb tracking system.

based on the described technology, to reconstruct tracking primitives in the upcoming Run 4. The target chosen for this upgrade is the reconstruction of segments in the scintillating fibres (SciFi) tracking stations, located downstream of the magnet ("T-tracks") (see Fig. 5). These tracks have an important role in LHCb tracking, being used both as seeds for the reconstruction of "Long" tracks (best-resolution tracks), and for the reconstruction of "Downstream" tracks (extending acceptance for long-lived particles); and even standalone. Their reconstruction is computationally heavy due to the large associated combinatorics, and in Run 2 it was only possible at the second trigger level. In the current run, it takes up a large fraction of the HLT1 computing power (∼350 NVIDIA A5000). Direct testing with simulated data showed that the average event processing time in HLT1 GPUs of 7.2 μs/event drops to 5.4 μs/event when T-track primitives are made available to jump-start the reconstruction. This figure includes all the additional decoding needed for the handling the extra raw data banks containing the T-track primitives. The 33% of HLT1 power freed in this way will then be available to significantly extend and improve the performance of HLT1 reconstruction and selection [8].

This project is described in detail in the Technical Design Report for enhancing the LHCb Data Acquisition system in LS3 [9], that at the time of this writing is under review by the LHC Experiments Committee.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We gratefully acknowledge the funding and support from the Istituto di Fisica Nucleare (Italy) through the CSN5-RETINA R&D project, and from the EU (NextGenerationEU) RRP through the Italian Ministero dell'Universitá e Ricerca (MUR), PRIN-2022 project 2022Z3K93E. We are also indebted to the LHCb Collaboration for its continued support and cooperation, particularly through its Online and Real Time Analysis projects.

References

[1] G. Bassi, et al., A FPGA-Based architecture for real-time cluster finding in the LHCb silicon pixel detector, IEEE Trans. Nucl. Sci. 70 (6) (2023) 1189.
[2] L. Ristori, An artificial retina for fast track finding, Nucl. Instrum. Methods A453 (1) (2000) 425–429.
[3] A. Abba, et al., A specialized processor for track reconstruction at the LHC crossing rate, JINST 9 (2014) C09001.
[4] F. Lazzari, et al., FPGA-based real-time data processing for accelerating reconstruction at LHCb, JINST 17 (04) (2022) C04011.
[5] F. Lazzari, et al., Demonstration of track reconstruction with FPGAs on live data at LHCb, EPJ Web Conf. 295 (2024) 02009.
[6] P. Hough, Machine analysis of Bubble Chamber Pictures, in: Proc. Int. Conf. High Energy Accelerators and Instrumentation, Vol. C590914, 1959.
[7] R. Aaij, et al., Evolution of the energy efficiency of LHCb's real-time processing, EPJ Web Conf. 251 (2021) 04009, http://dx.doi.org/10.1051/epjconf/202125104009, arXiv:2106.07701.
[8] W. Baldini, et al., Proposal for FPGA-based Tracking in the LHCb Downstream Region, Technical Report LHCb-PUB-2024-001, CERN, 2024.
[9] LHCb collaboration, LHCb Data Acquisition Enhancement TDR, Technical Report LHCB-TDR-025, CERN, Geneva, 2024.