PROCEEDINGS OF SCIENCE

Development of a High-Throughput Tracking Processor on FPGA boards

Riccardo Cenci*, **Federico Lazzari, Giovanni Punzi**
Università di Pisa & Istituto Nazionale di Fisica Nucleare - Sez. di Pisa, Pisa, Italy
E-mail: riccardo.cenci@cern.ch, federico.lazzari@cern.ch, giovanni.punzi@cern.ch

Pietro Marino†, **Michael J. Morello**
Scuola Normale Superiore & Istituto Nazionale di Fisica Nucleare - Sez. di Pisa, Pisa, Italy
E-mail: pietro.marino@cern.ch, michael.joseph.morello@cern.ch

Luciano F. Ristori
Fermi National Accelerator Laboratory, Batavia, IL, USA
E-mail: luciano@fnal.gov

Franco Spinella, Simone Stracka, John Walsh
Istituto Nazionale di Fisica Nucleare - Sez. di Pisa, Pisa, Italy
E-mail: franco.spinella@cern.ch, simone.stracka@cern.ch, john.walsh@cern.ch

We present the latest results on the prototype of a tracking processor capable of reconstructing events in a silicon-strip tracker at about 40 MHz event rate with sub-microsecond latency. The processor is based on an advanced pattern-recognition algorithm, called "artificial retina", inspired to the vision system of the mammals. We design and implement one of the first functional prototype of this processor on a DAQ board based on Alters Stratix III FPGAs. Then, in order to test the maximum rate capability, we port and optimize the processor on a high-speed board equipped with Altera Stratix V FPGAs. Future applications of this novel approach as real-time track trigger at LHC experiments are also discussed.

Topical Workshop on Electronics for Particle Physics
11 - 14 September 2017
Santa Cruz, California

*Speaker.
†Now at Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland.

© Copyright owned by the author(s) under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

PoS(TWEPP-17)136

1. The "Artificial Retina" Algorithm

Computing and storage demands of future LHC experiments at very high luminosity represent a challenge for HEP data processing, which calls for an efficient and scalable usage of the hardware. The increasing input rates and growing complexity of physics events, along with the finite bandwidth for writing to long term storage, call for sophisticated and computing intensive trigger algorithms. Moving part of the data processing, e.g. track reconstruction, to the online stage has multiple benefits: the stored event size can be reduced, trigger selection may be improved, and less processing has to be done offline. Given the steady increase of FPGA performances, their low power consumption and latencies, these devices are very suitable for implementing a tracking unit integrated in the DAQ architecture at a moderate cost, thus making event reconstruction primitives immediately available to event-building and high-level trigger (HLT) farms.

Our goal is to develop and implement a parallel computational methodology that allows to reconstruct events with an extremely high number (>100) of charged-particle tracks in silicon detectors at 40 MHz, thus matching the requirements for processing LHC events in real time. Our approach relies on a new pattern-matching methodology called "Artificial Retina" (AR) algorithm [1].

Inspired by the first stage of mammal vision, the AR algorithm is a highly-parallelized architecture able to reconstruct tracks at high speed. Its main features are the intelligent data distribution by a dedicated switching network and a continuous response of track patterns that allows interpolation. Consequently the internal bandwidth and number of needed patterns are significantly reduced. The mathematical aspects of the algorithm have some similarities with the "Hough transform" [2, 3], a method already applied for finding lines in image processing; however, the main challenge here is the design of the physical layout and the development of an implementation capable to sustain the event rate at high-luminosity LHC experiments [4]. More details about the algorithm implementation can be found in [5].

Compared to common DAQ and trigger systems, the total bandwidth increases significantly in the early stage, when hits are distributed by producing multiple copies, but greatly shrinks down later, when only the information about reconstructed tracks is kept. The parallel architecture allows to implement the system on multiple devices, but requires a full-mesh interconnection for a proper hits distribution. Therefore, its scalability depends on the number of high-speed I/O of each device. We find that the best device to implement a system based on the AR algorithm is FPGA. Compared to ASIC devices, FPGAs have higher flexibility and shorter time for designing and testing. To compare AR algorithm on FPGA with the computation of hits χ^2 on CPUs, we measure the event rate for both of them in the case of a very simple 3-layer tracker[1]. Measurements show that the event rate is better at lower occupancy for the CPU implementation, while at higher occupancy the AR algorithm is performing better, due to a different scaling of the two approaches.

2. The Functional Prototype

Given that the novelty of AR algorithm, first we design and implement a functional prototype [5]. The algorithm was configured to reconstruct tracks in a small tracker made by 6 single-

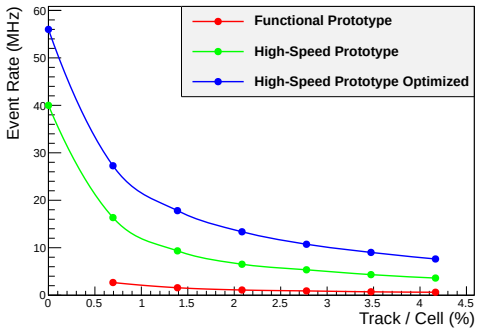
[1]For this tracker, the CPU implementation cannot be optimized further than computing the χ^2 for all the 3-hit combinations.

PoS(TWEPP-17)136

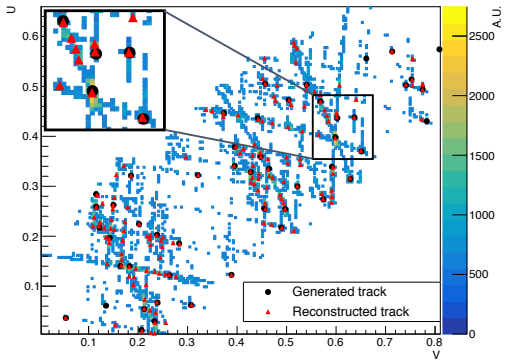
coordinate layers and no magnetic field. We parametrize the phase space of 2D tracks using the coordinates on the first and last layer, below referred to as u and v. The phase space is divided in cells, and each cell corresponds to a pattern. Each cell is implemented as a distinct processing unit, called engine, that accumulates weights proportional to the hit distance from the pattern track. Tracks are reconstructed looking for cells that are local maxima in the parameters phase space. Design and configuration of the system require a simulation of the tracker and of the data processing. Starting from the previous version of this C++ emulator [6], we largely extend its functionalities to include generation of configuration files for the electronic device, and of processed data at every stage for comparison with the real output. For this tracker we find that resolution and efficiency similar to offline tracking systems can be achieved using 3,000 cells.

We decide to implement the system in a multi-FPGAs board, the Tel62, developed for the DAQ of NA62 experiment [7]. The board is equipped with four Altera Stratix III chips for data processing. Given the limited flexibility of this board, we use one board for the switching network and another one for the cells logic, connected together through simple interfacing cards. The switching network is based on a simple dispatcher unit that can propagate any input data to any output, according to a preloaded LUT based on the hit coordinate and layer. In a single device we fit approximately 200 engines[2]. More details about the engine implementation can be found in [8].

We generate events using a toy simulation, load them into RAMs inside the chip, and feed then to the tracking processor in a continuous loop. The events are dispatched and processed correctly by the board running at the nominal clock rate (160 MHz). Because hits from each event are sent in sequence, the event rate is proportional to the tracker occupancy. We use events with an occupancy of from 1 to 6 tracks in the phase space covered by one chip, corresponding to few percent occupancy measured as number of tracks per cells. We measure an event rate of approximately 2.5 MHz down to 0.5 MHz, as shown in Fig. 1a, and a latency smaller than 1 µs.



(a) Event rate vs occupancy for the functional and high-speed prototypes described in the text. For the latter the rate is shown also for the configuration without multiple input lines.



(b) Values of accumulated weights for cells above threshold in the space of track parameters. Generated tracks and local maxima, equivalent to reconstructed tracks, are also overlap.

Figure 1

[2]The place and route tools from Quartus II report 90% of resources used on our specific device.

PoS(TWEPP-17)136

3. The High-Speed Prototype

After validating the algorithm using a functional prototype, we build another prototype using faster devices, in order to test the maximum performance achievable by the system. We port the design developed for the functional prototype to another board by DiniGroup, the DN0237. This board is equipped with two large Altera Stratix V FPGAs (almost 1M logic elements each), and much more flexible interconnections with respect to the Tel62. Due to the larger Stratix V FPGA, we implement in a single chip the logic previously fit in two paired Tel62 boards, thus using faster interconnections between the switching network and the engines. In addition, we optimize the design implementing multiple input lines to the same engine, because weights can be computed in parallel for different layers and summed at a later stage of the pipeline. After including the three improvements mentioned above (higher clock rate, faster interconnections, and multiple inputs), we reach an event rate approximately 20 times faster than the functional prototype, as shown in Fig. 1a. For this prototype we also measure a latency of less than 500 ns, perfectly in agreement with the requirements of DAQ systems planned for HL-LHC.

4. Application to a Real Case

After demonstrating high-speed performances, we apply our system to the LHCb Upgrade experiment in order to evaluate realistic performances and cost feasibility. This experiment will have no low-level trigger (LLT), and events will be sent directly to the HLT, implemented on a CPU farm, at the bunch-crossing rate (40 MHz) [9]. The HLT will be able to reconstruct only tracks originating inside the vertex tracker ("long" tracks). Reconstructing also tracks originating outside the vertex tracker ("downstream" tracks) will allow to increase the acceptance for long-lived particles (e.g. K_S^0, Λ's) by a factor 2-3. Anyway the reconstruction of "downstream" tracks is challenging because of the much-higher combinatorial in the tracking stations (axial and stereo strips) with respect to the vertex tracker (pixels). Given the Event Builder (EB) based on CPU, we design a tracking unit that can be integrated in the DAQ architecture using standard commercial PCIe cards equipped with FPGAs. We plan to obtain a copy of data from the readout, reconstruct the "downstream" tracks, and insert them in the EB network before the event is assembled. This approach, where tracks can be seen as the output of an additional "embedded track detector", has been included as proposal in the recent Expression of Interest presented to the LHCC [10].

We adapt the configuration described above to the tracking stations after the magnet, called SciFi. We expect about 50 reconstructable tracks in a SciFi quadrant[3]. Due to the number of cells fitting in a single FPGA and the maximum number of slots available for PCIe boards in the EB, we cover the track parameters space for each quadrant with approximately 20k cells. Using this configuration we process events with 50 tracks simulated with a simplified geometry ("toy"), and minimum-bias events fully simulated using the official LHCb simulation [11]. with the current geometry for LHCb Upgrade. The values of accumulated weights in cells above threshold for a typical "toy" event are shown in Fig. 1b. While 95% of the generated tracks are corresponding to local maxima, 48% of found maxima are false positives (called "ghosts"). The high number of "ghosts" is expected due to using only hits from axial layers and not from the stereo ones. When

[3]Quadrant can be considered independent for track reconstruction purpose.

PoS(TWEPP-17)136

using fully-simulated events that include effects from multiple scattering and residual magnetic field, we find that efficiency is above 90% for tracks with momentum greater that 3 GeV/c.

In the near future we plan to implement the SciFi configuration in the board to measure speed and latency with fully-simulated events, and later to integrate this design in the LHCb DAQ system. In parallel we want to add the information from stereo layers to reduce "ghosts", and from tracking station before the bending magnet to obtain a measure of track momentum.

5. Conclusions

We demonstrate that the AR algorithm is a feasible approach for a tracking trigger at low-level in a typical HL-LHC environment. We develop two prototypes of a system based on AR algorithm: a functional and a high-speed prototype. The former is aimed to demonstrate that the algorithm can be implemented on FPGA in a fully pipelined way and in less than 100 clock cycles. The latter focuses on exploring the highest speed reachable with current devices. For the high-speed prototype, we measure an event rate of about tenths of MHz, very close to the requirement, and latency less than 1 µs, much shorter than the time available before the event is built. Then, we start the development of a processor for reconstructing tracks originating outside the vertex detector at the LHCb Upgrade, showing performances comparable with offline software reconstruction when processing events fully simulated at the LHCb Upgrade conditions.

References

[1] L. Ristori, *An artificial retina for fast track finding*, *Nucl. Instrum. Meth.* **A453** (2000) 425–429.

[2] P. Hough, *Machine analysis of Bubble Chamber Pictures*, *Proc. Int. Conf. High Energy Accelerators and Instrumentation* **C590914** (1959) .

[3] P. Hough, *Method and mean for recognizing complex patterns*, *US Patent* **3069654** (1962) .

[4] W. H. Smith, *Trigger and Data Acquisition for hadron colliders at the Energy Frontier*, 1307.0706.

[5] R. Cenci, F. Bedeschi, P. Marino, M. Morello, D. Ninci, A. Piucci et al., *First Results of an "Artificial Retina" Processor Prototype*, *EPJ Web Conf.* **127** (2016) 00005.

[6] A. Abba, F. Bedeschi, M. Citterio, F. Caponio, A. Cusimano, A. Geraci et al., *A specialized track processor for the LHCb upgrade*, Tech. Rep. LHCb-PUB-2014-026, CERN, Geneva, Mar, 2014.

[7] B. Angelucci, E. Pedreschi, M. Sozzi and F. Spinella, *TEL62: an integrated trigger and data acquisition board*, *Journal of Instrumentation* **7** (2012) C02046.

[8] D. Ninci, *Real-time track reconstruction with FPGA at LHC*, Master's thesis, University of Pisa, Pisa, Italy, Dec, 2014.

[9] *LHCb Trigger and Online Upgrade Technical Design Report*, Tech. Rep. CERN-LHCC-2014-016. LHCB-TDR-016, May, 2014.

[10] LHCb collaboration, R. Aaij, B. Adeva, M. Adinolfi, Z. Ajaltouni, S. Akar, J. Albrecht et al., *Expression of Interest for a Phase-II LHCb Upgrade: Opportunities in flavour physics, and beyond, in the HL-LHC era*, Tech. Rep. CERN-LHCC-2017-003, CERN, Geneva, Feb, 2017.

[11] M. Clemencic, G. Corti, S. Easo, C. R. Jones, S. Miglioranzi, M. Pappagallo et al., *The LHCb Simulation Application, Gauss: Design, Evolution and Experience*, *Journal of Physics: Conference Series* **331** (2011) 032023.

PoS(TWEPP-17)136