FERMILAB-PUB-16-184-CMS
ACCEPTED

An "artificial retina" processor for track reconstruction at the full LHC crossing rate

A. Abba[d,f], F. Bedeschi[a], F. Caponio[d,f], R. Cenci[a,c,*], M. Citterio[d], A. Cusimano[d,f], J. Fu[d], A. Geraci[d,f], M. Grizzuti[d,f], N. Lusardi[d,f], P. Marino[a,c], M. J. Morello[a,c], N. Neri[d], D. Ninci[a,b], M. Petruzzo[d,e], A. Piucci[a,b], G. Punzi[a,b], L. Ristori[g], F. Spinella[a], S. Stracka[a,b], D. Tonelli[h], J. Walsh[a]

[a] *Istituto Nazionale di Fisica Nucleare - Sez. di Pisa, Pisa, Italy*
[b] *Università di Pisa, Pisa, Italy*
[c] *Scuola Normale Superiore, Pisa, Italy*
[d] *Istituto Nazionale di Fisica Nucleare - Sez. di Milano, Milano, Italy*
[e] *Università di Milano, Milano, Italy*
[f] *Politecnico di Milano, Milano, Italy*
[g] *Fermi National Accelerator Laboratory, Batavia, Illinois, USA*
[h] *Cern, Geneva, Switzerland*

Abstract

We present the latest results of an R&D study for a specialized processor capable of reconstructing, in a silicon pixel detector, high-quality tracks from high-energy collision events at 40 MHz. The processor applies a highly parallel pattern-recognition algorithm inspired to quick detection of edges in mammals visual cortex. After a detailed study of a real-detector application, demonstrating that online reconstruction of offline-quality tracks is feasible at 40 MHz with sub-microsecond latency, we are implementing a prototype using common high-bandwidth FPGA devices.

Keywords: Pattern Recognition, Trigger; Real-time Online Tracking
PACS: 29.85.Ca

1. Introduction

The increase in energy and luminosity forecast at LHC in the next decade creates a serious challenge for selecting interesting events at the LHC experiments. The discriminating power of usual signatures, such as the high transverse momentum of leptons or the high transverse missing energy, is greatly reduced due to the large number of interactions for bunch crossing (pile-up). Reconstructing tracks in the event in real time would provide a very efficient way to trigger events, but performing such a task at the LHC crossing rate is problematic due to large combinatorial and size of the associated data flow, other than requiring a massive computing power. For accomplish this at a reasonable cost we plan to use a new pattern-recognition algorithm called "artificial retina" [1], and inspired to the quick detection of edges in mammals visual cortex.

2. The Artificial "Retina" Algorithm

Experimental studies of the mechanism of vision in the mammals have shown that, for the first stage, neurons are tuned to recognize a specific shape on specific region of the retina, called receptive field. Those neurons receive signals only from that retina region, in order to reduce the number of connections and the data amount. The neuron response to a stimulus is proportional to how close the shape of the stimulus is to the shape for which the neuron is tuned to. Generated in parallel, the responses of neurons are then interpolated to create a preview of the image edges in about 30 ms, corresponding to about 30 clock cycles. Those concepts of early vision can be used for a track reconstruction task, as described below.

In the first phase, performed offline by a common PC, we create our algorithm configuration that consists of two mappings. The parameters that describe uniquely a track are two if we are in a 2D projection and no magnetic field, or up to five if we are in a 3D space and a magnetic field is present. The space of track parameters is divided into cells, which mimic the receptive fields of the retina. The center of each cell identifies a track in the detector that intersects the layers in spatial points called receptors. The first mapping connects each cell with those receptors. Considering a group of contiguous cells, where the track parameters vary by a small amount, the corresponding receptors in the detector layers would belong to a limited area. The second mapping connects clusters of cells to areas of the detector and this information is recorded in a LUT.

The second phase, that runs in real-time on high-speed devices, has three steps. In step 1, detector hits are distributed only to a reduced number of cells according the LUT's created in the first phase. For each incoming hit, in step 2 the algorithm accumulates in each cell a Gaussian weight w proportional to the distance with the receptor computed as follows:

$$w = \exp\left(-\frac{d_l^2}{2\sigma}\right) \quad (1)$$

where d_l is the distance, on the layer l, between the hit and the

*Corresponding author
Email address: riccardo.cenci@pi.infn.it (R. Cenci)

Operated by Fermi Research Alliance, LLC under Contract No. De-AC02-07CH11359 with the United States Department of Energy.

corresponding receptor, and σ is a parameter of the algorithm, that it can be adjusted to optimize the sharpness of the response of the receptors. After all hits are processed, in step 3 tracks are identified as local maxima of weights sum, above a certain threshold, over the cells grid. For a track resolution similar to offline algorithm, the cells grid does not require a high granularity, because significant better resolution on track parameters can be easily obtained computing the centroid of the sums for the cells surrounding each maximum.

3. Study for a Real Detector Application

To evaluate the performances and the robustness of the algorithm in a real and complex HEP detector, we perform a detailed study using as benchmark the LCHb tracker designed for the 2020 upgrade [2]. We arbitrarily chose to parametrize tracks with five parameters u, v, d, z_0, and k defined in [2]. We develop a detailed C++ simulation of the algorithm that can be interfaced with the official LHCb Monte Carlo simulation [3] to compute the receptors, and that can process directly LHCb MC events simulated with 2020 conditions. To evaluate the efficiency we consider only "reconstructable" tracks in the acceptance of our system, requiring at least three hits on VELO layers and two hits on UT layers, and applying cuts on momentum ($p > 3\,\mathrm{GeV}/c$) and transverse momentum ($p_T > 200\,\mathrm{MeV}/c$). Using a grid with around 20k cells, we obtain reconstruction and resolution performances at the same level of offline processing. Basic design and HDL implementation on Altera Stratix V FPGA shows that the event processing time is about 150 clock cycles, that corresponds to a latency of 0.5 μs using a clock at 350 MHz. All the cells can be fit into 32 FPGA devices.

4. First Prototype

To demonstrate the feasibility of a track processing system based on the artificial retina algorithm we design a first prototype based on a 6 single-coordinate layers tracker equivalent to an independent sector of the current IT detector in LHCb. The algorithm for this tracker was configured using a C++ simulation similar to the one described above, and tested using simulated and real data from LHCb [4]. In this case a track can be parametrized using only two quantities, and we choose the first and last layer coordinates. We implement the algorithm logic on the FPGA Altera Stratix III, and test the system using the Tel62 board [5]. The algorithm simulation shows that good reconstruction efficiency can be obtained with 3k cells, that can be fit into 8 boards, corresponding to 32 chips. As shown in Fig. 1, step 1 is implemented into one board, then hits data are transferred to another board for performing step 2 and 3.

4.1. Technical details

The Tel62 board was developed by INFN-Pisa for the DAQ of NA62 experiment [5]. This board includes 4 Stratix III chip for data processing, connected through a high-speed link (10 Gbit/s) to a master chip (another Stratix III) that collects the processed data and controls the other FPGA's. The main clock

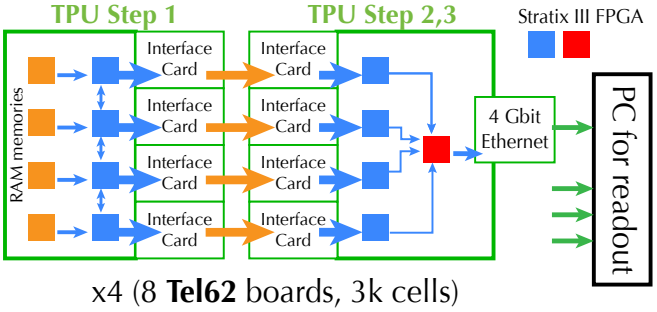

Figure 1: Architecture of the first prototype using Tel62 boards.

is 40 MHz, while the processing internal to the FPGA's is done mostly at 160 MHz. Each processing chip is connected to a 2 Gbyte DDR2 RAM through a high-speed bus (40 Gbit/s), to a mezzanine connector that can host various interface cards for I/O (5 Gbit/s), and to the neighbour FPGA's (2.5 Gbit/s). The master FPGA is also connected to a mezzanine card with Ethernet ports (2.5 Gbit/s). A slow-control interface, implemented on an embedded PC that sits on the board, can be used to monitor all the devices and to perform various standalone tests.

4.2. Results

We implement and simulate step 2 and 3 of the algorithm inside the Stratix III chip [6]. Loading the design on the real board, we are able to run the system at 160 MHz and to receive the processed data on the readout PC with no error. Assuming detector occupancy from real data and using the switch simulation, we compute that the prototype can substain a maximum event rate of 1.8 MHz with a latency smaller than 1 μs.

5. Conclusions

We report on the first implementation of the "artificial retina" algorithm for track reconstruction. The first prototype is currently under development and results show that the track processing unit based on this algorithm is properly working on the real device, and satisfies the latency and performance requirements. We plan to achieve higher throughtput rate using larger and faster devices already available on the market, where we can process multiple events at the same time.

References

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

[2] A. Abba *et al.*, A specialized track processor for the LHCb upgrade, CERNa-LHCb-PUB-2014-026.

[3] M. Clemencic *et al.*, The LHCb simulation application, Gauss: Design, evolution and experience, J. Phys. Conf. Ser. 331, 032023 (2011).

[4] A. Piucci, Reconstruction of tracks in real time at high luminosity environment at LHC, Master thesis, https://etd.adm.unipi.it/theses/available/etd-06242014-055001/.

[5] F. Spinella *et al.*, The TEL62: A real-time board for the NA62 Trigger and Data Acquisition. Data flow and firmware design, IEEE Nucl. Sci. Symp. Conf. Rec., 1 (2014).

[6] D. Ninci, Real-time track reconstruction with FPGA at LHC, Master Thesis, https://etd.adm.unipi.it/theses/available/etd-11302014-212637/.