PROCEEDINGS OF SCIENCE

Testbeam results for the first real-time tracking system based on artificial retina algorithm

A.Abba[*], **F.Caponio**[*], **S.Coelli, M.Citterio, J.Fu, A.Merli, M.Monti, N.Neri, M.Petruzzo**[†]
Università degli Studi di Milano and INFN-Milano, Milano
E-mail: marco.petruzzo@mi.infn.it

We present the testbeam results of the first embedded tracking system prototype based on artificial retina algorithm, capable to reconstruct tracks in real time with a latency $< 1\mu$s and with track parameter resolutions comparable with the offline results. The tracking system is based on silicon strip telescope and a custom DAQ board based on commercial FPGA, in which the artificial retina algorithm has been implemented. The prototype has been tested with 180GeV/c protons at a maximum trigger rate of 280kHz. The testbeam has been carried out at CERN SPS and the obtained results for track parameters show good agreement with offline results and with the simulated response of the system.

The 25th International workshop on vertex detectors
September 26-30, 2016
La Biodola, Isola d'Elba, ITALY

[*] now at Nuclear Instruments srls.
[†] Speaker.

PoS(Vertex 2016)062

1. Introduction

The INFN-Retina project aims at developing a fast track finding system prototype capable to operate at 40MHz event rate with hundreds of track per event, for the high-luminosity LHC experiments. The reconstruction of charged particle trajectories at high energy physics experiments is a non trivial task and requires to identify candidate trajectories between multiple possibilities of combinations of hits in the detector. Fast track finder devices exploit the pattern recognition in parallel using custom processors that compare the measured hits to precomputed track patterns, then the track parameters are obtained via simplified algorithms based on the constraints given by the pattern matching. An example of fast track finder is the *Silicon Vertex Trigger* (SVT) that was used in the CDF experiment [1]. It was based on associative memories to perform the pattern recognition, then a linearized fitting algorithm was applied using fast FPGAs. Results were available with a latency of about 10μs at an event rate of 30kHz, with good quality track parameters. A similar concept is applied in the ATLAS *Fast TracKer* (FTK) [2].

In order to exploit the full potential of the high luminosity phase of LHC new detectors and new trigger schemes are needed. In particular including the tracking information in the first stage of the trigger chain will allow to reduce the data rate while maintaining a good efficiency and purity of the signal. Nevertheless both the need to reconstruct the events at 40MHz and the event complexity itself make real-time tracking at LHC experiments extremely challenging and new approaches are necessary to deal with this problem.

2. Artificial retina algorithm

The artificial retina algorithm is an innovative approach for fast track finding that is inspired from the mechanism of visual receptive fields in the visual cortex [3]. The algorithm is based on a grid of cellular units covering the available space of track parameters and tuned to recognize specific tracks. In particular each cellular units evaluate the response ("Weight") to the measured positions from the tracking device that is proportional to how close the precomputed track is to the measured points. All the cellular units evaluate the Weight function in parallel and a track candidate is identified for each local maximum in the grid. The track parameters are then obtained via interpolation of the Weight values near the local maxima. Thanks to the high level of parallelization of the algorithm, it is particularly suitable for implementation in FPGA and application to high energy physics experiments

2D artificial retina algorithm We describe here the application of the artificial retina algorithm for a case study of a 2D tracking system, as represented by the silicon telescope we used for the presented tests.

We consider a track defined in the (z,x) 2-dimensional space, where z is the axis of the telescope and x is the measured position in each plane. We define (z_f,x_f) and (z_l,x_l) the coordinates of the intersection of a track in the first and last tracking plane, respectively, and we define the track parameters as $x_\pm = (x_f \pm x_l)/2$ and the auxiliary variables $z_\pm = (z_f \pm z_l)/2$. Using this notation, a track trajectory is defined as

$$x(z) = x_+ + x_- \, (z - z_+)/z_- \, . \tag{2.1}$$

The cellular units are uniformly distributed in the space of track parameters (x_-, x_+) and each one is labeled by the indexes (i, j) in the grid and is associated to a set of track receptors positioned at the intercepts of the ideal track with the tracking planes as shown in Fig.1. The response of

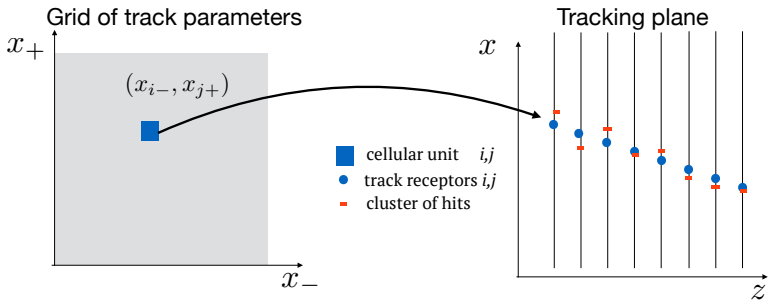

Figure 1: Example of association of a cellular unit in the grid of track parameters (left) with its corresponding track receptors in the tracking plane (right).

a cellular unit to a measured hit depends on the distance s_{ijk} between the k-th measured cluster and the track receptor in the tracking plane associated to cell (x_{i-}, x_{j+}). In particular the receptor provides a Gaussian response evaluated as

$$W_{ijk} = \begin{cases} \exp\left(-\frac{s_{ijk}^2}{2\sigma^2}\right) & \text{if } |s_{ijk}| < 2\sigma \\ 0 & \text{otherwise} \end{cases} , \tag{2.2}$$

where $s_{ijk} = x_k - x_{j+} - x_{i-}\ (z - z_+)/z_-$ and σ is the width of the Gaussian response. The latter parameter needs to be tuned for optimal response and a good choice is $\sigma \simeq \Delta$, where Δ represents the granularity of the grid.

For each measured cluster, there is bundle of lines whose intercept is compatible with that point. This means that in the space of track parameters the cluster (z_k, x_k) produces an excitation on a subset of cellular units placed along a line defined by

$$x_+ = x_k - x_- \frac{z_k - z_+}{z_-} . \tag{2.3}$$

The total weight function is defined as the sum of the responses as

$$W_{ij} = \sum_k W_{ijk} , \tag{2.4}$$

and clusters belonging to the same track produce a local maximum as shown in Fig.2.
A candidate tracks is identified from a local maximum and the associated track parameters are obtained by Gaussian interpolation of the weight function along the x_-, x_+ axes, for the cell with maximum and the neighbours.
The results of the Gaussian interpolations are given by

$$\begin{aligned} x_{-,\text{rec}} &= x_{-i} + \frac{\Delta}{2} \frac{\ln(W_{i-1\ j}/W_{ij}) - \ln(W_{i+1\ j}/W_{ij})}{\ln(W_{i-1\ j}/W_{ij}) + \ln(W_{i+1\ j}/W_{ij})} , \\ x_{+,\text{rec}} &= x_{+j} + \frac{\Delta}{2} \frac{\ln(W_{i\ j-1}/W_{ij}) - \ln(W_{i\ j+1}/W_{ij})}{\ln(W_{i\ j-1}/W_{ij}) + \ln(W_{i\ j+1}/W_{ij})} . \end{aligned} \tag{2.5}$$

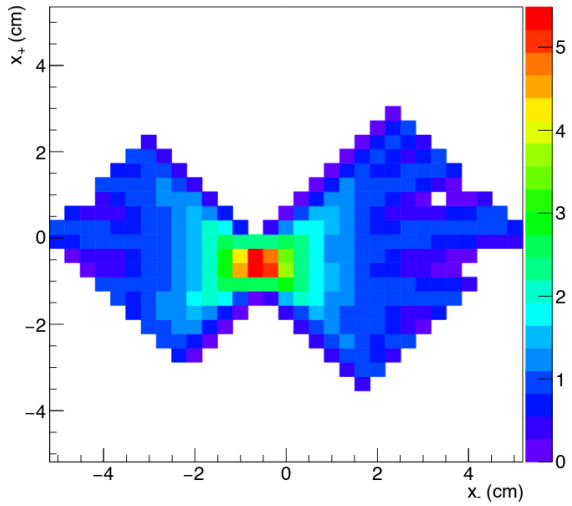

Figure 2: Response of the artificial retina for an identified track from real test-beam data. The excitations from the track hits produce a local maximum.

Artificial retina architecture The artificial retina algorithm has been implemented in FPGA. The scheme of the architecture is represented in Fig.3 and it is organized in three main blocks:

- *switch*, delivers in parallel the hits from the DAQ to the regions of cellular engines where the hits are expected to produce a non negligible response;
- *engines*, evaluate the calculation of the responses associated to the cellular units. The evaluation is performed in parallel for all the engines receiving an hit;
- *track fitter*, evaluates the parameter of the reconstructed track by interpolation of the weight value of the local maximum and its neighbour cells.

In particular the switch consists in a 4×16 dispatcher, each input receives the data from 2 tracking planes while the 16 outputs are connected to 16 different regions of cellular units, dividing the space of track parameters in a 4×4 coarse grid. Each region hosts 32 engines, for a total of 512 covering the whole space. The switch consists in a network of two-way sorters with programmable look-up tables (LUT). A two-way sorter is module with 2inputs/2outputs that receive the hits and forward them to zero, one or both the outputs according to the comparison of the hit information with the LUT. An *hold logic* is implemented to take buffer the data when multiple hits try to access the input ports, or the output ports are not available due to an hold given by the following two-way sorter, engine or the track fitter.
An engine correspond to a cellular unit. It receives the hit data from the switch and performs the evaluation of the response in four pipelined stages: the calculation of the signed distance s_{ijk}, its absolute value $\left|s_{ijk}\right|$, the Gaussian field response to the single hit and finally the sum of this value to the previous excitations (from other hits). The evaluation of the retina response is stopped when an *EndOfEvent* signal come from the DAQ: at this point the engines forward the information about weight values to track fitter, then the value is reset and they are ready for the next event.
The track fitter is composed by 16 interpolation units connected to blocks of 32 engines. If one or

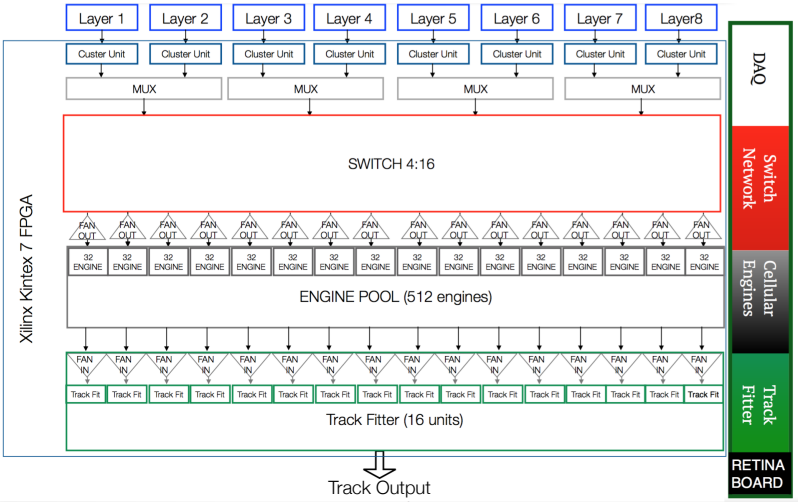

Figure 3: Detailed view of the architecture of the artificial retina implemented in FPGA for the tracking system prototype.

more engines of the block are identified as local maxima the track parameters (x_-, x_+) of the reconstructed tracks are obtained via two parabolic interpolations (instead of Gaussian interpolations) of the weight values along both x_-, x_+ axes and the obtained values are then corrected using LUTs. The use of the parabolic interpolation is motivated by the higher quantity of resources needed to perform the evaluation of logarithms defined in Eq. (2.5).

3. Tracking system prototype

A silicon strip telescope together with a custom DAQ board, the artificial retina, have been developed for the real-time tracking system prototype. Details are given in the following paragraphs.

Silicon strip telescope The telescope consists of 8 planes of single-sided silicon strip sensors (STM OB2) and it has been build and assembled at INFN-Milano. Each sensor has 512 channels with 183μm strip pitch with an active area of about 100cm^2 and 500μm thickness. The readout of the sensors is performed using *TTHybrids* from LHCb experiment. Each TTHybrid hosts 4 Beetle chips[4] and provide the analog readout of a full sensor. The telescope is equipped with a linear and a rotation stage in order to test the telescope in different positions and orientations with respect to the beam. The trigger is provided from the coincidence of two plastic scintillators placed before and after the tracking planes. In Fig.4 a picture of the telescope, mounted on beam at the CERN SPS is shown. In this configuration there are two symmetrical arms made of 3 sensors, with a sensor in the middle. The distance between the planes within each arm is 4cm while the distance between the central sensor and the nearest of each arm is 8cm.

DAQ board with artificial retina algorithm A custom DAQ board, called MAMBA board, based on Xilinx Kintex 7 (XC7K410T) FPGA has been designed for the readout of the telescope and the implementation of the artificial retina algorithm. It can read out to 32 Beetle chips and a single board can read out the entire telescope. It is equipped with 12-bit ADCs for the digitization

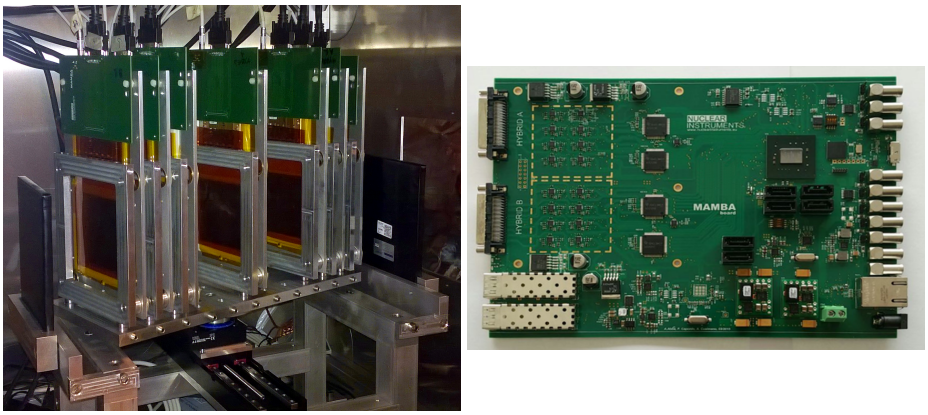

Figure 4: Silicon strip telescope mounted at the testbeam at CERN SPS (left) and MAMBA DAQ board (right).

of the analog signals from the sensor modules and a programmable threshold is applied to the strip signals to identify the hits. Clusters of hits from adjacent hits are made by a cluster unit and sent to the artificial retina for the track reconstruction. The track parameters are output from the track fitter and stored to disk via a USB 3.0 interface, together with the values of the weight function for debugging purposes. The Beetle chip sampling rate is 40MHz while the maximum accepted trigger rate is limited at about 280kHz due to the time needed to read out the strip signals stored in the Beetle analogue pipeline.
The artificial retina algorithm, instead, works at a clock frequency of 150MHz and a clock cycle corresponds to 6.7ns Results are provided with a latency $< 1\mu$s, according to number of clock cycles required to perform the track reconstruction. The amount of FPGA resources needed for the implementation of the DAQ and the artificial retina are 3% and 68%, respectively. Table 1 shows the contributions to the latency and the percentage of resources needed by the three blocks of the artificial retina.

Module	Clock cycles	FPGA resources(%)
Switch	14	7
Engines	12	37
Track Fitter	68	24
Total	94	68

Table 1: Latency of the retina response in number clock cycles of the FPGA and percentage of logic resources allocated for each individual module.

4. Testbeam results

The full chain of the tracking system (sensors, DAQ, artificial retina) has been successfully tested using 180GeV/c protons during at a testbeam at CERN SPS. Tracks have been reconstructed in real time using the artificial retina. The track parameters are in good agreement with the results obtained from a simple χ^2-minimization algorithm. For debugging purposes the retina response

to real testbeam data has been simulated and reproduces the results obtained from the artificial retina algorithm running in FPGA. The resolution on the track parameters obtained by the online algorithm has been evaluated comparing retina track parameters with the track parameters from the offline (χ^2-minimization) reconstruction.
The distribution of the residuals for (x_-, x_+) are shown in Fig.5 and have been fitted with a Gaussian function. The obtained widths are $\sigma_{x_-} = 12.5\mu$m and $\sigma_{x_+} = 14.9\mu$m .

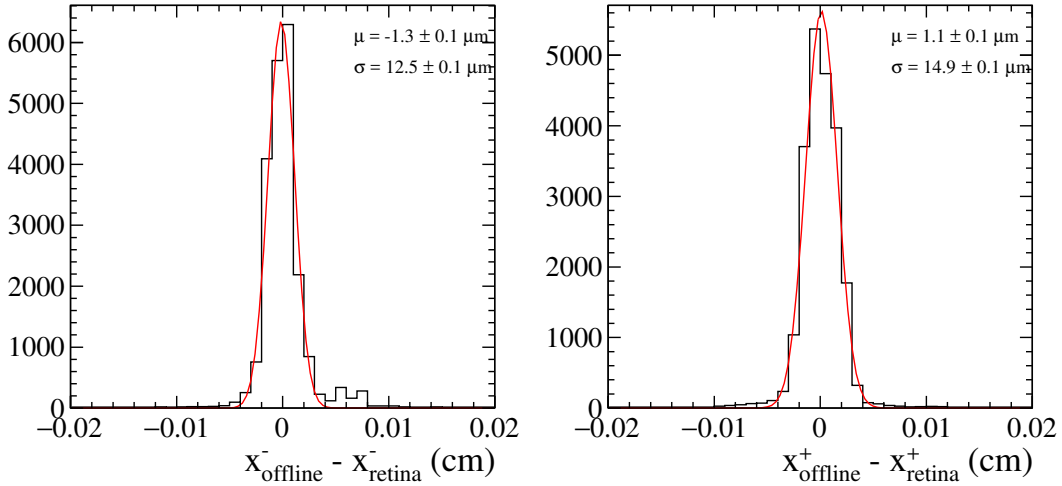

Figure 5: Distribution of the residuals for the track parameters evaluated using a simple χ^2 minimization algorithm (offline) and track parameters from the artificial retina algorithm

References

[1] Ashmanskas B, Barchiesi A, Bardi A, Bari M, Baumgart M, Belforte S, et al. The CDF Silicon Vertex Trigger. Nucl Instrum Meth 2004;A518:532-6.

[2] Shochet M, Tompkins L, Cavaliere V, Giannetti P, Annovi A, Volpi G. Fast TracKer (FTK) Technical Design Report. Tech. Rep. CERN-LHCC-2013-007. ATLAS-TDR-021; CERN; Geneva; 2013.

[3] Ristori L. An artificial retina for fast track finding. Nucl Instrum Meth 2000;A453:425-9.

[4] Lochner S, Schmelling M. The Beetle Reference Manual - chip version 1.3, 1.4 and 1.5. Tech. Rep. LHCb-2005-105. CERN-LHCb-2005-105; CERN; Geneva; 2006.