Performance of a high-throughput tracking processor implemented on Stratix-V FPGA

R. Cenci[a,b], F. Lazzari[a,b,*], P. Marino[b,c], M. J. Morello[b,c], G. Punzi[a,b], L. F. Ristori[d], F. Spinella[b], S. Stracka[b], J. Walsh[b]

[a] *Università di Pisa, Pisa, Italy*
[b] *Istituto Nazionale di Fisica Nucleare - Sez. di Pisa, Pisa, Italy*
[c] *Scuola Normale Superiore, Pisa, Italy*
[d] *Fermi National Accelerator Laboratory, Batavia, IL, USA*

Abstract

Two of the biggest challenges for future HEP experiments at hadron colliders are triggering and track reconstruction of high-multiplicity events, where collisions have multiple primary vertices. The highly-non-linear scaling of computing power required for these tasks encourages the adoption of non-traditional, specialized architectures. Amongst them, the "artificial retina" approach, inspired by the architecture of the vision system in the living brain, promises large efficiency of hardware utilization, low-power and low-latency when implemented in state-of-art FPGA devices.

The INFN-RETINA project has been a 3-year effort dedicated to investigate the potential of that approach in a real-time tracking processor at Level-0 of the LHC HEP experiments. We present results from studies performed on a prototype system capable of carrying out track reconstruction in a generic 6-layer silicon-strip detector with sub-μs latency at an event rate in excess of 30 MHz, and how a large-scale system can be implemented on multiple boards interconnected with high-speed optical links. Possible applications to real experimental environments will be also discussed.

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

1. Introduction

Computing and storage demands of future LHC experiments at very high luminosity represent a challenge for HEP data processing. Events with a high-number of primary vertices require sophisticated and computing intensive trigger algorithms, mainly based on the reconstruction of high-quality tracks. The highly-non-linear scaling of computing power required for these tasks encourages the adoption of non-traditional, specialized architectures.

Given the steady increase of FPGA performances, and 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 was to develop and implement a system that allows to reconstruct high-quality tracks at LHC beam crossing rate (30 MHz). Our approach relies on a new pattern-recognition methodology called "Artificial Retina" (AR) architecture [1].

2. The "Artificial Retina" architecture

Inspired by the first stage of mammal vision, the AR architecture is a highly-parallelized architecture able to reconstruct tracks at high speed. The mathematical aspects of the architecture are related to the "Hough transform" [2, 3], a method already applied for finding lines in image processing; however the main challenge is the design and the implementation of a system capable to sustain the event rate at high-luminosity LHC experiments (30 MHz).

The tracks parameters space is mapped into a matrix of cells. A custom switch delivers hits only to appropriate cells reducing the input bandwidth per cell. For each cell, the corresponding logic (referred to below as Engine) performs a weighted sum of the distances between the hits and the track trajectory. Parameters of reconstructed tracks are obtained interpolating the responses of nearby cells. Engines work in a fully parallel way. More details about the architecture implementation can be found in [4, 5].

The AR architecture is highly parallelized: this allows to implement the system on multiple devices, but requires a mesh interconnection for a proper hits distribution. Therefore, its scalability depends on the number of high-speed I/O of each device.

3. The INFN-RETINA project

The INFN-Retina project aims to demonstrate the feasibility of a system based on AR architecture able to reconstruct tracks at rates expected at Level-0 of the LHC HEP experiments. Two fundamental steps of the project are: verify that the system is capable to sustain an event rate of 30 MHz with a latency lower

*Corresponding author
Email address: lazzarifederico@gmail.com (F. Lazzari)

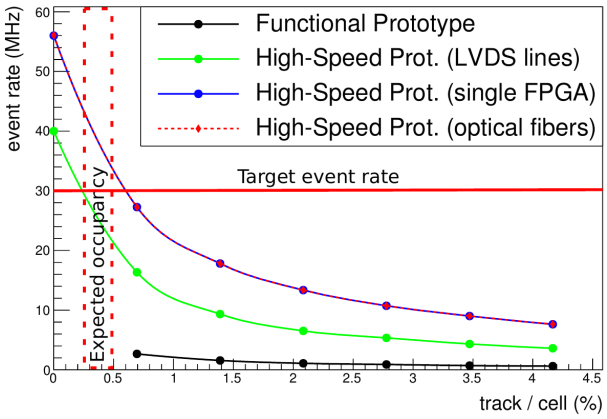

Figure 1: Event rate vs occupancy for configurations described in the text.

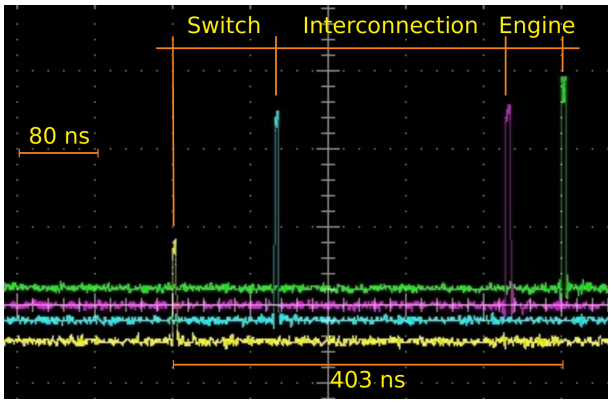

Figure 2: Latency measurement for the high-speed prototype using the optical fibers interconnection.

than few μs, and that it can be implemented on more interconnected chips without reducing the throughput.

To demonstrate the system feasibility we design two prototypes based on a generic 6 single-coordinate layers tracker, with no magnetic field. We choose as track parameters the track coordinates on first and last layer. Such a tracker is equivalent to the LHCb SciFi tracker [6]. Typically SciFi is crossed by 200 tracks per event. We define the occupancy of the system as the ratio between the number of tracks on the detector and the number of the cells of the AR. Considering that we can map the track parameters space with 80k cells using <150 FPGAs [7], the occupancy is 0.25%.

3.1. Implementation

We implement a high-speed prototype using a commercial board with two Altera Stratix V. The board exposes also all the high-speed serial lines for each FPGA.

We implement three different configurations on this board: to compare the performances of the system on Stratix V with respect to the first prototype of AR architecture [5], we implement the switch into one FPGA and the Engines into the second one, the hits data are transferred through LVDS lines; for measuring the maximum reachable event rate, we optimized the design and implemented the entire system in a single FPGA; to demonstrate the possibility of multi-board implementation, we split the optimized design in the two FPGAs and we connected them with optical fibers.

4. Results

The event rate is proportional to the tracker occupancy. We test the system using events with an occupancy between 0 and 4.2%. For the high-speed prototype implemented in a single FPGA, we measure an event rate >30 MHz when the occupancy is <0.5% [7], reaching the project goal. Fig. 1 show event rate measurements for different configurations.

Connecting the switch and the Engines through optical fibers, we achieve the same event rate, demonstrating that this system can be distributed on more FPGAs without decreasing the throughput. In this configuration we measure a latency of $\approx$400 ns (Fig. 2), where the largest contribution is coming from the optical interconnection. This measurement demonstrates that even a distributed system satisfies the latency requirements.

A tracking processor based on this concept has been included in the recent LHCb Expression of Interest for future upgrades [9].

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] A. Abba et al., A specialized track processor for the LHCb upgrade, Tech. Rep. LHCb-PUB-2014-026, CERN, Geneva, Mar, 2014.
[5] R. Cenci et al., First Results of an "Artificial Retina" Processor Prototype, EPJ Web Conf. 127 (2016) 00005.
[6] LHCb Collaboration, LHCb Tracker Upgrade Technical Design Report, CERN-LHCC-2014-001; LHCB-TDR-015.
[7] R. Cenci et al., Development of a High-Throughput Tracking Processor on FPGA Boards, PoS(TWEPP-17) 136.
[8] 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).
[9] LHCb collaboration, 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.