Fast algorithm for real-time rings reconstruction.

R. Ammendola^{1},*M. Bauce*^{2,3,4}, *A. Biagioni*^{2}, *S.Capuani*^{3,5}, *S. Chiozzi*^{6,7}, *A. Cotta Ramusino*^{6,7}, *G.Di Domenico*^{6,7}, *R. Fantechi*^{2,8}, *M. Fiorini*^{6,7}, *S.Giagu*^{2,3}, *A. Gianoli*^{6,7}, *E. Graverini*^{8,9}, *G. Lamanna*^{8,10,*}, *A. Lonardo*^{2}, *A. Messina*^{2,3}, *I. Neri*^{6,7}, *M. Palombo*^{5,11,12}, *F. Pantaleo*^{9,10}, *P. S. Paolucci*^{2}, *R. Piandani*^{8,9}, *L. Pontisso*^{8}, *M.Rescigno*^{2}, *F. Simula*^{2}, *M. Sozzi*^{8,9}, *P. Vicini*^{2}

[1]INFN Sezione di Roma “Tor Vergata”, Via della Ricerca Scientifica 1, 00133 Roma, Italy
[2]INFN Sezione di Roma“La Sapienza”, P.le A. Moro 2, 00185 Roma, Italy
[3]University of Rome “La Sapienza”, P.le A.Moro 2, 00185 Roma, Italy
[4] CERN, Geneve, Switzerland
[5]IPCF-UOS Roma, Physics Department, ”Sapienza” University of Rome, Rome, Italy
[6]INFN Sezione di Ferrara, Via Saragat 1, 44122 Ferrara, Italy
[7]University of Ferrara, Via Saragat 1, 44122 Ferrara, Italy
[8]INFN Sezione di Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy
[9]University of Pisa, Largo B. Pontecorvo 3, 56127 Pisa, Italy
[10] INFN Sezione di Frascati, Italy
[11]CEA/DSV/I2BM/MIRCen, Fontenay-aux-Roses, France
[12]CEA-CNRS URA 2210, Fontenay-aux-Roses, France, France
* Corresponding author. E-mail: gianluca.lamanna@cern.ch

DOI: http://dx.doi.org/10.3204/DESY-PROC-2014-05/42

The GAP project is dedicated to study the application of GPU in several contexts in which real-time response is important to take decisions. The definition of real-time depends on the application under study, ranging from answer time of μs up to several hours in case of very computing intensive task. During this conference we presented our work in low level triggers [1] [2] and high level triggers [3] in high energy physics experiments, and specific application for nuclear magnetic resonance (NMR) [4] [5] and cone-beam CT [6]. Apart from the study of dedicated solution to decrease the latency due to data transport and preparation, the computing algorithms play an essential role in any GPU application. In this contribution, we show an original algorithm developed for triggers application, to accelerate the ring reconstruction in RICH detector when it is not possible to have seeds for reconstruction from external trackers.

1 Introduction

Low level trigger RICH reconstruction is very challenging. But the possibility to use in trigger decisions the information coming from Cherenkov rings, would be very useful in building stringent conditions for data selection. Standard algorithms are difficult to use in an on-line environment: some of them requires an external seed, some of them is not adequate in terms of resolution and noise immunity, all of them are too slow. In particular, in high intensity experiments high data rate requires fast data processing setting the maximum complexity allowed for the on-line reconstruction algorithms. The recent development in high throughput processors provides new opportunities in using information coming from rings in RICH detectors directly in the first trigger level. The GPUs (Graphics Processing Units), in particular, are designed with a parallel architecture in order to exploit the huge computing power offered by a big number of computing cores working together on the same problem. This kind of structure (the so called SIMD (Single Instruction Multiple Data)) allows the design of new parallel algorithms suitable for pattern recognition and ring fitting, with very high computing throughput and relatively small latency. In this paper we will discuss a new algorithm heavily based on the possibility to run concurrently several computing threads on the same processor. The GPU implementation of this algorithm and the results in terms of performances will be discussed.

2 Requests for an on-line ring reconstruction

Most of standard single ring algorithms [7] can not be trivially used in real-time application. The geometrical fits, like for instance Gauss-Newton, Chernov-Lesort and others, are very accurate, but most of them require an initial seed to start, while the algebraic fits, like for instance Taubin, Crawford and others, are less accurate, but usually faster. In both cases it is not easy to extend the ring reconstruction to a multi-ring problem. Specialized algorithms, like Hough transform, fitQun, APfit, based on likelihood and similar, are usually slow with respect to the requirements imposed by high intensity experiments. A good algorithm candidate to be used in low level trigger applications should have the following characteristics:

- seedless,
- multi-Ring reconstruction,
- fast enough to cope with event rate of tens of MHz,
- accurate (events reconstruction with offline resolution).

In addition, depending on the specific experiment requirements, noise immunity and stability of the fitting time with respect to the complexity should be required. Presently no suitable algorithms are available on the market.

3 A physics case: the NA62 RICH

As physics case for the study described in this paper, we consider the NA62 RICH detector. The NA62 experiment at CERN [8] has the goal of measuring the branching ratio of the ultra-rare decay of the charged kaon into a pion and a neutrino anti-neutrino pair. The main interest in this decay is linked to the high precision theoretical prediction of its branching ratio, at the level of few percent, since it is almost free of non-parametric theoretical uncertainties. Because of this, a precise measurement with a sample of 100 events would be a stringent test of the Standard Model, being also highly sensitive to new physics. The trigger is a key system to obtain such a result. In its standard implementation, the FPGAs on the readout boards of each sub-detector participating to the L0 trigger, compute simple trigger primitives. The maximum latency allowed for the synchronous L0 trigger is related to the maximum data storage available on the data acquisition boards; its value was chosen to be rather large in NA62 (up to 1 ms). Such a time budget allows in principle the use of more complex but slower trigger implementations at this level. This would have the benefit of increasing the trigger selectivity for the $K^+ \to \pi^+\nu\bar{\nu}$ process, and would allow the design of new trigger condition to collect additional physics processes.

The NA62 RICH [9] is a 17 meters long, 4 meters in diameter, 1 atm Neon filled detector, used to distinguish pions and muons in the 15-35 GeV/c range. The Cherenkov light is focused on two spots equipped with about 1000 small (16 mm in diameter) phototubes each. The ring maximum radius is about 20 cm, while the average number of hits is 20 per ring. Typical NA62 events are one charged particle in final state ($K^+ \to \pi^+\pi^0$ and $K^+ \to \mu^+\nu$), but interesting three tracks events are also possible ($K^+ \to \pi^+\pi^+\pi^-$, Lepton Flavor Violation modes, etc.).

4 The Almagest

Most of the algorithms mentioned above are efficient for single rings. A strategy to implement a multi-rings reconstruction should use a two step process: first collect the hits (pattern recognition) from the same circle and then perform the fit. Obviously the pattern recognition is the most time consuming and difficult part. We present a new idea based on elementary geometry to address this problem.

In his treatise on astronomy, the Almagest, Ptolemy derived several geometrical results. In particular, in Euclidean geometry, one of Ptolemy's theorems connects the lengths of the four sides and the two diagonals of a cyclic quadrilateral, that is quadrilateral whose vertexes all lie on a single circle. *In a convex quadrilateral, if the sum of the products of its two pairs of opposite sides is equal to the product of its diagonals, then the quadrilateral can be inscribed in a circle.* In the particular case shown in figure 1, the theorem states that:

$$\overline{AC} \cdot \overline{BD} = \overline{AB} \cdot \overline{CD} + \overline{BC} \cdot \overline{DA}$$

This is an useful condition for our problem, since it does not require any other information but the distances between hits inside the RICH. Using this theorem it is possible to decide if a point should be considered for a single ring fit or not. Assuming that the maximum number of hits coming from the NA62 RICH is 64, checking all the combinations of all the hits in groups of four (more than 600000), to decide if the points belong to the same ring, is unfeasible in on-line reconstruction.

To reduce the number of tests, one possibility is to choose few triplets, i.e. a set of three hits assumed to belong to a single ring, and iterate through all the other hits while checking whether Ptolemy's relation is satisfied. In this way, in the worst case scenario of a 64 hit NA62 event, the number of iterations for each triplet turns out to be 61. The hits collected in this phase will be used to fit the ring with a classical single-fit algorithm (like the Taubin method). The problem converges only if the points of the starting triplet come from the same ring. Obviously the choice of the test triplets is decisive for the efficiency of the algorithm. Considering events with two or three rings like, for example, the one shown in figure 2, to maximize the probability that all the points of a triplet belong to the same ring we decided to use 8 triplets and in particular the ones obtained with three points from the top, bottom, right, left and from the U and V diagonal directions.

A key point of this approach is the possibility to run more than one "triplet" at the same time. Modern processors offer the possibility to vectorialize the algorithms in order to exploit parallel computing. In particular the so called streaming processors have the SIMD (Single Instruction Multiple Data) architecture that allow to run the same program on different sets

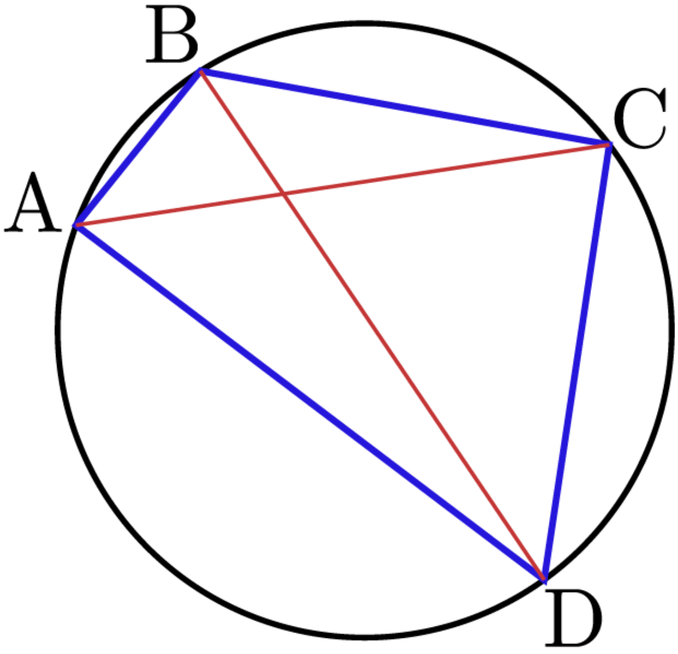

Figure 1: Ptolemy's Theorem.

of data concurrently. Among SIMD processors, the GPUs (Graphics Processing Units) offer a huge computing power, as will be discussed in the next section.

In figure 3 a preliminary result on the reconstruction efficiency for the multi-ring selection is shown. The rings are obtained using GEANT4 Montecarlo, including electromagnetic showers, delta rays and hadronic interaction. For this reason, additional spurious hits are possible. The efficiency strongly depends on hits distribution: asymmetric hits distribution for a ring, for instance, could introduce a bias in the triplets selection described above, as well as points too close each other could affect the single ring (Taubin) fit. Further cuts on hits distance and position can improve the efficiency significantly. The average time to reconstruct a multi-ring event is about 1 μs, using one single GPU.

5 The GPU

Graphic processors represent a viable way to fill the gap between a multi-level system, with a hardware-based lowest level trigger, and a system which would not require preliminary real-time hardware processing on partial event information. Indeed GPUs do provide large raw computing power even on a single device, thus allowing to take complex decisions within a latency that can match significant event rates. This computing power is given by the different processor architecture with respect to the standard CPUs. In GPU more transistors are devoted to computing with respect to the CPU, the number of cores in a GPU can be easily of the order

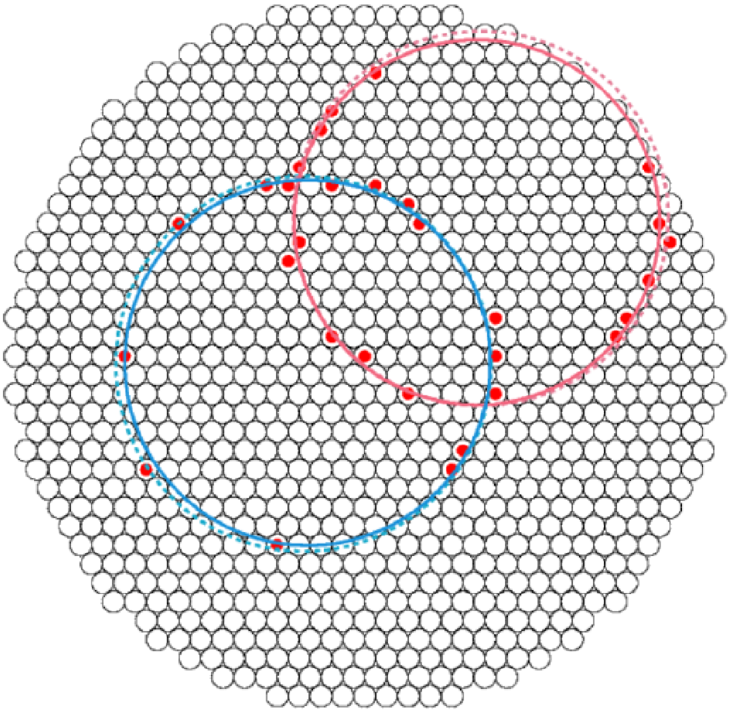

Figure 2: Two rings reconstructed by Almagest's algorithm (solid line). The dotted ring represent the real ring position. The small circles represent the PMTs positions within the NA62 RICH.

of few thousands, with respect to the few units in multi-cores CPU. In addition the computing dedicated structure of the GPU is obtained by reducing the cache dimension, increasing the memory bus and simplifying the control stage.

In recent times GPUs are used for General Purpose computing (GPGPU), and not only for graphics applications. In the High Performance Computing sector there are two major GPU vendors: AMD and NVIDIA. For the moment we are concentrating on NVIDIA, because they can be programmed at a lower level allowing better control of the hardware through the software. GPU parallelism, on Graphic cards produced by NVIDIA, is exposed for general-purpose computing thanks to an architecture called Compute Unified Device Architecture (CUDA). The CUDA architecture is built around a scalable array of multi-threaded Streaming Multiprocessors (SMs). A SM is designed to execute hundreds of threads concurrently. Each thread is execute on a single CUDA core. The CUDA cores share a very fast on-chip memory, to allow data exchange and temporary caching. The computing power of recent GPUs can easily exceed 4 Teraflops.

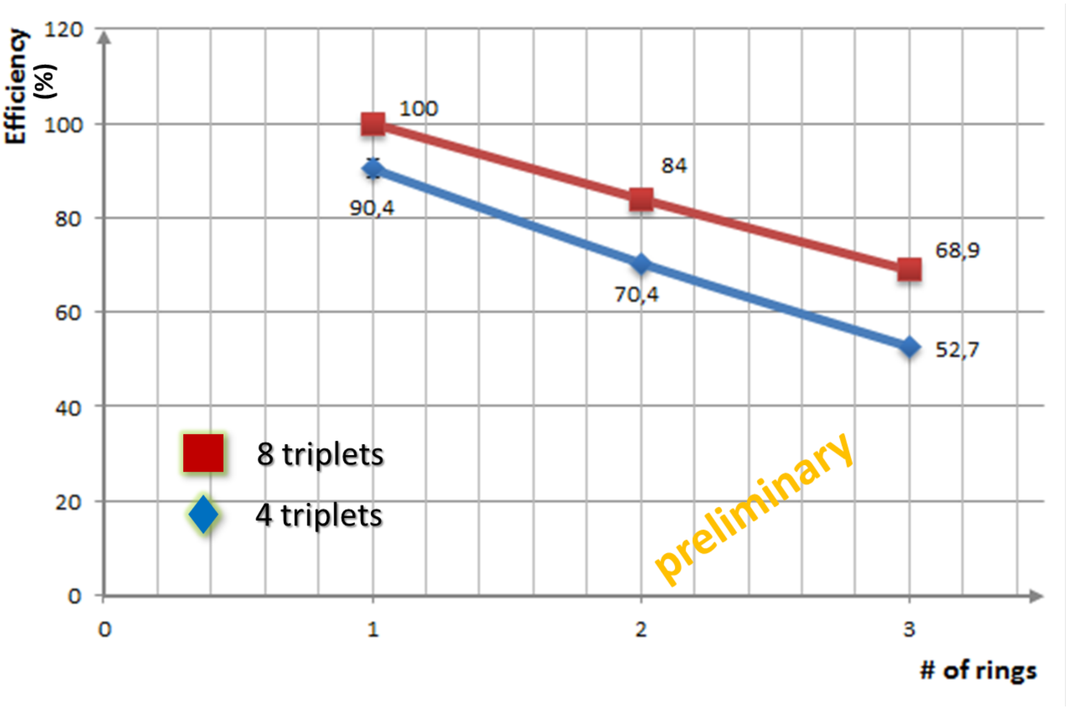

Figure 3: Preliminary efficiency of the algorithm as a function of the number of rings, for 4 triplets (horizontal and vertical) and 8 triplets (horizontal, vertical and diagonals) version. As expected, the efficiency increases with the number of triplets considered.

6 GPU in real-time

The use of GPUs in lowest trigger levels requires a careful assessment of their real-time performance. A low total processing latency and its stability in time (on the scales of hard real-time[1]) are indeed requirements which are not of paramount importance in the applications for which GPUs have been originally developed. This issue is related to the fact that in common usage of GPUs as graphics co-processors in computers, data are transferred to the GPU - and results are transferred back - through the PCIExpress computer bus. Moreover, in order to better exploit the parallel architecture of GPUs, their computing cores should be saturated, thus requiring the computation on a significant number of events after a buffering stage.

GAP (GPU Application Project) is focused on the use of GPU in real-time, both in High Energy Physics and medical imaging. We use two techniques to reduce the data transportation latency: PFRING and NANET. The first is a new fast capture driver, designed to avoid any redundant operation on the data received in a standard NIC (Network Interface Card)[10], while the second is a board named NANET [11], based on an FPGA, in which the possibility to copy directly the data from the FPGA to the GPU is implemented. Preliminary results, on single ring search kernel [12], shows similar performance (figure 4 and figure 5). In the case of

[1] A system is defined as “hard” real-time if a task that temporally exceed its deadline causes irrecovarable damage to the system.

PFRING we used a TESLA K20 board, while for the NANET test we used, a little bit older, TESLA M2070. The K20 is almost a factor of two faster than the M2070 in executing the kernel, but all the other components of the latency are almost the same (both boards are used as X16 PCIExpress gen.2). The main component of the total latency is the gathering time needed to collect a sufficient number of events to better exploit the GPU computing power and to hide the overhead latency due to data transmission. Times of the order of 150/200 μs per event is the limit of the present technology. The latency time fluctuation is well below the μs. This is very important in using GPU for synchronous low level trigger. The behaviour of the GPUs, with respect to the CPUs, is almost deterministic (the non-deterministic component is due to the GPU-CPU interaction in data transfer and kernel loading) and very stable.

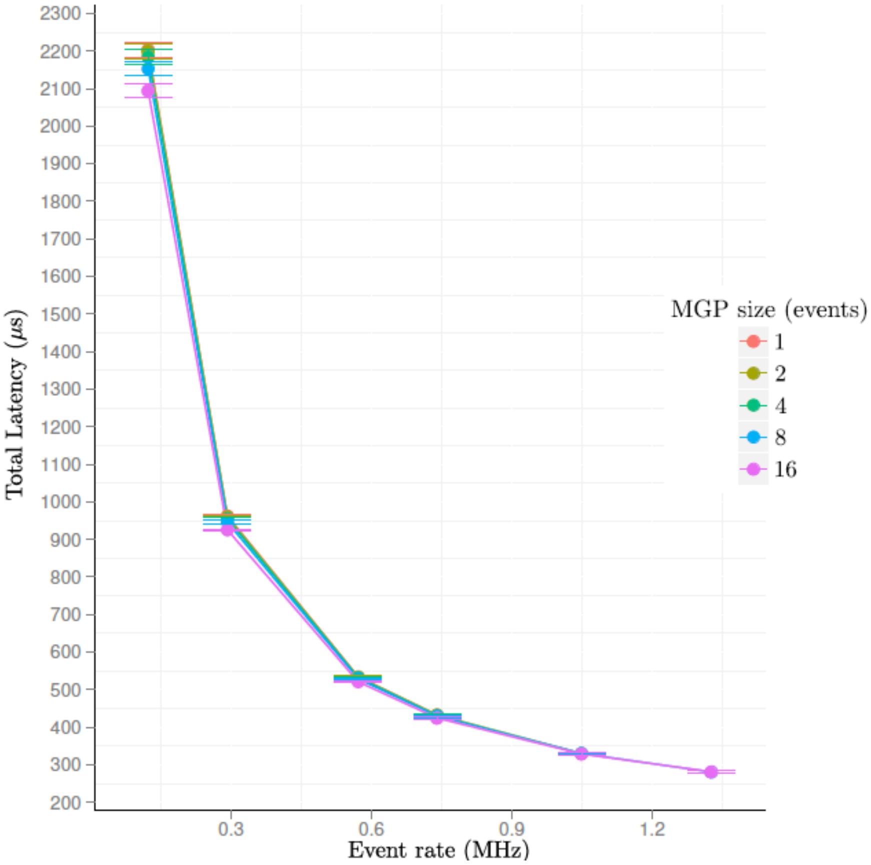

Figure 4: Total time (including data transfer and computing for a single ring kernel) as a function of the rate, for a buffer of 256 events, for PFRING driver solution. The time is independent on the MGP (the number of events per data packet from the detector).

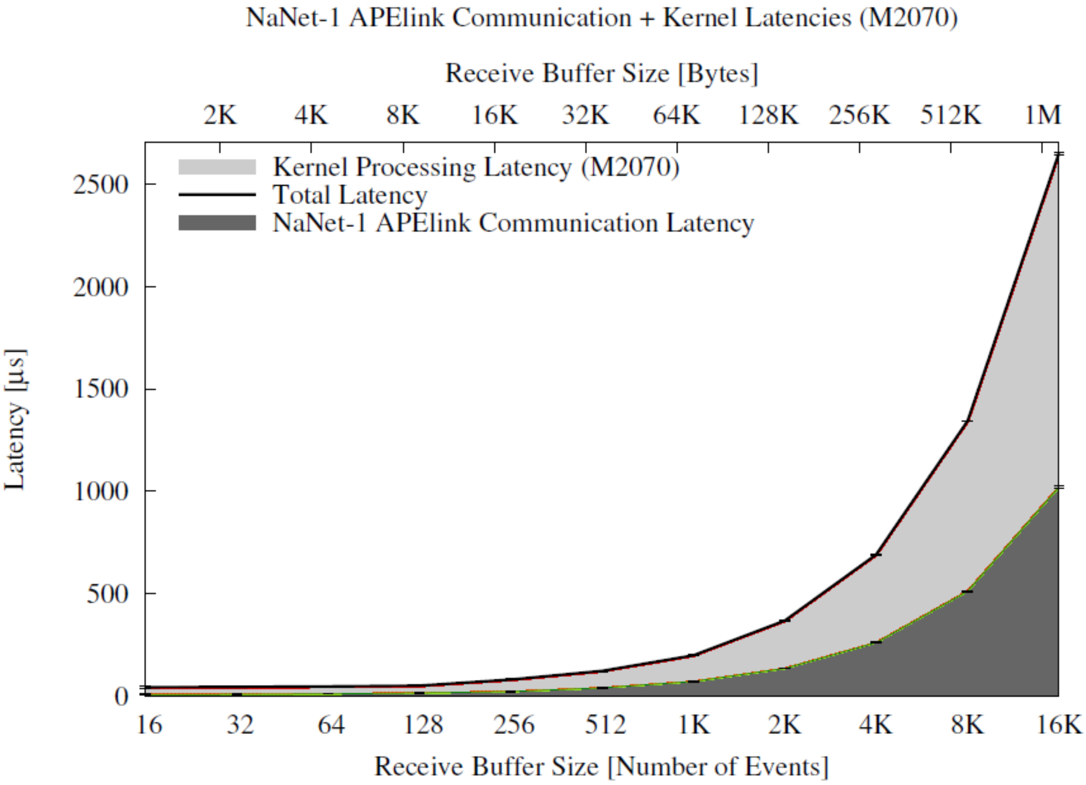

Figure 5: Total time (including data transfer and computing for a single ring kernel) as a function of the buffer dimension, for NANET board solution.

7 Conclusions

We have presented a new algorithm for ring reconstruction. This algorithm, based on Ptolemy's theorem on quadrilaterals, is suitable to be used for circular pattern recognition. A parallel implementation of this algorithm is proposed to run on a GPU, the standard video processor. The computing power for this kind of device is huge for parallel application. Preliminary tests, using the NA62 RICH as physics case, are very encouraging. The maximum latency is in the order of 150/200 μs, while the computing time is 50 ns for single ring, and 1 μs for multi-rings events. Rare events search experiment, like NA62, would benefit on the possibility to exploit on-line ring reconstruction to define selective trigger for data acquisition. A "demonstrator" is in preparation for testing, in parassitic way, during the next NA62 run.

Acknowledgment

The GAP project is partially supported by MIUR under grant RBFR12JF2Z "Futuro in ricerca 2012".

References

[1] M. Fiorini et al. *GPUs for the realtime low-level trigger of the NA62 experiment at CERN.* GPU in HEP 2014 conference.

[2] A. Lonardo et al. *A FPGA-based Network Interface Card with GPUDirect enabling realtime GPU computing in HEP experiments.* GPU in HEP 2014 conference.

[3] M. Bauce et al. *The GAP project: GPU applications for High Level Trigger and Medical Imaging* GPU in HEP 2014 conference.

[4] M. Palombo et al. *Estimation of GPU acceleration in NMR medical imaging reconstruction for realtime applications.* GPU in HEP 2014 conference.

[5] M. Palombo et al. *GPU-parallelized model fitting for realtime non- Gaussian diffusion NMR parametric imaging.* GPU in HEP 2014 conference.

[6] G. Di Domenico et al. *Fast Cone-beam CT reconstruction using GPU* GPU in HEP 2014 conference.

[7] N. Chernov *Circular and linear regression: Fitting circles and lines by least squares* Monographs on Statistics and Applied Probability (Chapman & Hall/CRC), Volume 117 (2010).

[8] NA62 Collaboration, *NA62: Technical Design Document, https://cds.cern.ch/record/1404985*, 2010.

[9] B. Angelucci*et al.*, Nucl. Instrum. Meth. A **621** (2010) 205. (*see also M.Piccini at the same conference RICH2013*)

[10] *http://www.ntop.org*

[11] R. Ammendola, A. Biagioni, O. Frezza, F. Lo Cicero, A. Lonardo, P. S. Paolucci, D. Rossetti and F. Simula *et al.*, J. Phys. Conf. Ser. **396** (2012) 042059.

[12] G. Collazuol, G. Lamanna, J. Pinzino and M. S. Sozzi, Nucl. Instrum. Meth. A **662** (2012) 49.