Real-Time Use of GPUs in NA62 Experiment

Gianmaria Collazuol, *Università di Padova*, Vincenzo Innocente, *CERN*, Gianluca Lamanna, *CERN*, Felice Pantaleo, *CERN*, and Marco Sozzi, *Università di Pisa*

Abstract—We describe a pilot project for the use of GPUs in a real-time triggering application in the early trigger stages at the CERN NA62 experiment, and the results of the first field tests together with a prototype data acquisition (DAQ) system.

This pilot project within NA62 aims at integrating GPUs into the central L0 trigger processor, and also to use them as fast online processors for computing trigger primitives. Several TDC-equipped sub-detectors with sub-nanosecond time resolution will participate in the first-level NA62 trigger (L0), fully integrated with the data-acquisition system, to reduce the readout rate of all sub-detectors to 1 MHz, using multiplicity information asynchronously computed over time frames of a few ns, both for positive sub-detectors and for vetos. The online use of GPUs would allow the computation of more complex trigger primitives already at this first trigger level.

We describe the architectures of the proposed systems, focusing on measuring the performance (both throughput and latency) of various approaches meant to solve these high energy physics problems.

The challenges and the prospects of this promising idea are discussed.

Index Terms—GPU, Real-time, Pattern recognition, Trigger, CUDA, RICH, NA62.

I. INTRODUCTION

THE use of accelerators and in particular of Graphic Processing Units (GPUs) in High Performance Scientific Computing has grown a lot in the last half decade bringing desktop and laptop computers to the Terascale (i.e. bringing computational power beyond a Teraflop), clusters to the Petascale and, in the foreseeable future, supercomputers to the Exascale.

GPUs are made of massively parallel SIMD Streaming Multiprocessors (SM) with a three-level memory hierarchy, each one with a different access time. The architectural design of the GPUs is pushed by the fast-growing game market that requests more and more FLOPS per video frame for use in games. This is the reason why GPU design devotes a bigger chip area to floating point calculation optimized for the execution of a massive number of threads in parallel.

In the field of high energy physics, several groups are pursuing the use of GPUs for Maximum Likelihood fit [1] inside the RooFit [2] data analysis package or for Monte Carlo simulations of particle interactions [3] [4] inside the Geant4 Framework [5].

The use of GPUs presented here is different: we aim at employing GPUs for taking real-time decisions in a trigger system for the NA62 experiment, studying both its use in a high level software trigger and in a fixed-latency hardware trigger. As a first example of possible applications of GPUs, we discuss the 10 MHz pattern matching and ring fitting in the NA62 RICH (see [6] for more details). For what concerns

the high level trigger, the ALICE collaboration at CERN has developed a software trigger that exploits commercial GPUs for the track fitting [7].

II. THE NA62 EXPERIMENT

In this section we will briefly introduce the NA62 experiment, focusing in particular on the trigger system. Further details on the experiment can be found in [8]. The NA62 experiment aims at measuring $O(100)$ $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ events in two years of data taking. The Standard Model prediction for the Branching Ratio (BR) of this decay mode is $(8.7 \pm 0.7) \times 10^{-11}$, with a theoretical irreducible error of a few percent. For this reason $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ is a “golden mode” to test the CKM structure of the Standard Model and an excellent probe of new physics beyond the Standard Model, in a complementary way with respect to the direct search.

From an experimental point of view, the measurement is very challenging due to the smallness of the BR and the presence of a large background, mainly from $K^+ \rightarrow \pi^+ \pi^0$ and $K^+ \rightarrow \mu^+ \nu$ (with BR of 20.7% and 63.5% respectively). The best current measurement of this decay is based on 7 candidates collected by the E787+E949 Brookhaven experiments [9] leading to a value of $BR = (1.74_{-0.89}^{+1.3}) \times 10^{-10}$. The biggest difference with respect to these experiments is that NA62 will exploit the kaons decay in flight technique using an unseparated hadron beam (kaon $\sim 6\%$) of 75 GeV/c, produced from 400 GeV/c protons from the CERN SPS impinging on a fixed beryllium target. The decay region will be housed in a $\sim 120m$ long $\sim 2.5m$ diameter vacuum tube, in order to reduce the secondary interactions of both the decay products and the primary beam.

In Figure 1 the layout of the experiment is shown. The background rejection will be based both on high resolution kinematics reconstruction and on veto systems. The former will be achieved by measuring with high precision both the outgoing pion momentum (STRAWS tracker) and the incoming kaon momentum (GIGATRACKER), while the photon veto rejection will be done using rings made of lead glass blocks (LAV) for the photons with a large angle, an electromagnetic calorimeter (LKr) for the photons in the forward direction and other small calorimeters (IRC and SAC) for the photons close to the beam line. The particle identification system (CEDAR, RICH and MUV) will help in rejecting the non-kinematically constrained component of the total background.

The RICH, in particular, must identify pions and muons in the momentum range 15 GeV/c to 35 GeV/c, giving a μ suppression factor better than 10^{-2} with a good time resolution. Čerenkov light is produced in a 18 m long, 3.7 m wide tube filled with neon at atmospheric pressure. The light is

Fig. 1. Layout of the NA62 experiment.

reflected by a composite mirror of 17 m focal length, focused on two separated spots. The two spots are equipped with ~ 1000 PMs, each being 1.8 cm in diameter. After amplification and discrimination [10], the PM signal is digitized by high resolution TDCs. A typical pion ring, for averaged accepted momentum, is identified with ~ 20 firing PMs, as predicted by Monte Carlo and confirmed with a full-length prototype [11]. The time resolution was measured to be better than 100 ps for all momenta in the considered range.

A. The NA62 Trigger system

In order to collect $O(100)$ Standard Model events, an intense beam and a reliable data acquisition and trigger system (TDAQ) are needed. An efficient on-line selection of candidates represents an important issue for this experiment because of the large reduction to be applied on data before tape recording. On the other hand, a loss-less data acquisition system is mandatory to avoid adding artificial detector inefficiencies when vetoing background particles; this last requirement is less common in standard readout and trigger systems. For the above reasons the detectors and the DAQ systems are integrated in a completely unified digital system. In order to reduce the event rate from 10 MHz to tens of kHz, the TDAQ is structured in a three level system. The first level (L0) will be completely hardware based while the other levels (L1 and L2) will be based on software: the L1 decision is taken on single-subdetector reconstructed quantities, while the L2 decision is taken on the fully reconstructed event with high resolution. The L0 trigger primitives are constructed in the same board (TEL62) in which the data are stored to wait for the trigger decision. The TEL62 board is a general purpose board (an upgraded version of the TELL1 board developed by EPFL for the LHCb experiment [12]) with 5 FPGA and large buffers to store the data, waiting for the trigger decision delivered to the board through the CERN standard TTC interface. On the TEL62, up to 4 daughter boards can be mounted. For the majority of the detectors in NA62, time is the most important information to provide. Therefore a daughter board with 4 HPTDC [13] has been designed, in order to have 512 channels, each with 100 ps time resolution, in a single TEL62 board. The trigger primitives produced in the TEL62 are sent to the L0 trigger processor (LOTP) using an Ethernet connection with low level protocol. The reduction factor of 10, at L0, will be obtained using positive information from RICH and CHOD, combined with the veto information from LKr, MUV and LAV. The L0 trigger decision is broadcast through TTC to all the TEL62 with a fixed latency of 1 ms. The interesting events will

go into the L1 via an Ethernet connection. The L1 will apply, in software, a further reduction factor of 10. The decision at this level is based on the data coming from a single detector, in particular STRAW or RICH. In the L2 the full event is reconstructed and a more complete event selection is applied in order to reduce the rate to tens of kHz. The latency in the software levels is not defined, but all the events have to be processed before the next accelerator burst (order of 20-30 s).

III. FAST RING RECONSTRUCTION FOR THE NA62 RICH DETECTOR

As a first example of the use of GPUs in the NA62 trigger system, we have studied the possibility to reconstruct rings in the RICH. The center and the radius of the Čerenkov rings in the detector are related to the angle and the velocity of the particle. This information can be employed at the trigger level to increase the purity and the rejection power for many triggers of interest. The ring reconstruction could be useful both at L0 and L1. In both cases, because of the high rate of 10 and 1 MHz respectively, the computing power required is significant. The GPUs can offer a simple solution of the problem. The use of GPUs in the L1 is straightforward: the GPU can be used to offload the computation. On the other hand the L0 is a small latency synchronous level, and the possibility to use the GPU must be verified.

In order to test feasibility and performance, as a starting point we have implemented the algorithm described by Crawford [14], for single ring finding in a sparse matrix of 1000 points (centered on the PMs in the RICH spot) with 20 firing PMs (“hits”) on average. For the moment we focused on single ring recognition.

A. Crawford Method

Consider a circle of radius R , centered in (x_0, y_0) and a list of points (x_i, y_i) . Following the method described by Crawford[14], the following relations exist:

$$x_0^2 + y_0^2 - R^2 = \frac{1}{N} \{ 2x_0 \sum x_i + 2y_0 \sum y_i - \sum x_i^2 - \sum y_i^2 \}. \quad (1)$$

$$x_0 \{ \sum x_i^2 - \frac{(\sum x_i)^2}{N} \} + y_0 \{ \sum x_i y_i - \frac{\sum x_i \sum y_i}{N} \} = \frac{1}{2} \{ \sum x_i^3 + \sum x_i y_i^2 - \sum x_i \frac{\sum x_i^2 + \sum y_i^2}{N} \}, \quad (2)$$

and

$$x_0 \{ \sum x_i y_i^2 - \frac{\sum x_i \sum y_i^2}{N} \} + y_0 \{ \sum y_i^3 - \frac{\sum y_i^2 \sum y_i}{N} \} = \frac{1}{2} \{ \sum x_i^2 y_i + \sum y_i^3 - \sum y_i \frac{\sum x_i^2 + \sum y_i^2}{N} \}. \quad (3)$$

Equations (2) and (3) can easily be solved for x_0 and y_0 . Eventually R can be determined using equation (1).

IV. ALGORITHM DESCRIPTION

The evaluation of the coordinates of the center and the radius of the circle for a single event proceeds as follows:

- the average of the coordinates are computed:

$$x_m = \frac{\sum x_i}{N}, y_m = \frac{\sum y_i}{N};$$

- the differences between each photomultiplier coordinates and x_m and y_m are evaluated giving respectively u_i and v_i ;
- for each photomultiplier the algorithm computes: $u_i^2, v_i^2, u_i v_i, u_i^3, v_i^3, u_i^2 v_i, u_i v_i^2$;
- the sums $\sum u_i^2, \sum v_i^2, \sum u_i v_i, \sum u_i^3, \sum v_i^3, \sum u_i^2 v_i, \sum u_i v_i^2$ are evaluated;
- eventually the coordinates of the center and the radius of the circle can be evaluated using equations (1), (2) and (3).

A. Parallelization Strategy

Since an event rate of $\sim 600\text{MB/s}$ is expected, a coarse grained parallelism is straightforward. However event parallelism is not the best solution to exploit all the computational potential of a massively parallel architecture like a GPU.

For this reason a reduction kernel has been developed. This kernel splits the original array $a[2N]$ in two sub-arrays $b[N]$ and $c[N]$, and evaluates in parallel in shared memory the partial sums $a[i]+b[i]$ for all the elements in the sub-arrays. This procedure is repeated until the results array contains only one element that is the total sum of the elements of the original array.

Brent's theorem suggests to spawn $O(N/\log N)$ threads and that each does $O(\log N)$ sequential work to avoid to have too many threads idling. Since the time complexity is $O(\log N)$, all the threads will cooperate for $O(\log N)$ steps giving a new total cost of $O(N)$.

Furthermore the evaluation of the arrays $u_i^2, v_i^2, u_i v_i, u_i^3, v_i^3, u_i^2 v_i, u_i v_i^2$ perfectly suits the SIMD paradigm.

Another level of parallelism is represented by the concurrency of data transfers and kernel executions on NVIDIA devices of compute capability greater than 2.0. In fact, using streams, it is possible to overlap the kernel execution with the data transfer of the results of the previous packet back to host memory and with the data transfer of new data belonging to the next packet in queue to the GPU memory. Hence a good fraction of the latency due to the data transfer can be hidden.

B. Implementation

An event comes as an array of structures (x_i, y_i) , where x_i and y_i are the spatial coordinates of the photomultiplier that has been fired. These arrays of structures, for their sparse access pattern to the memory, do not represent the best structure the GPU can compute on. These array of structures, by mean of a simple transposition, become a structure of arrays $(x_0, \dots, x_{n-1}), (y_0, \dots, y_{n-1})$ in order to achieve higher bandwidth and throughput. This allows us to achieve a speed-up in performance of $\sim 60\%$.

Fig. 2. Throughput for a variable number of events inside each packet.

To limit the overhead connected to the data transfers from and to the GPU, it's not possible to send the events on the fly even if this is not optimal for the latency. A packet of events will be sent to the GPU memory only when a good size has been reached, or if the maximum latency could be exceeded.

Since `cudaMalloc`, `cudaMallocHost`, `cudaFree` and `cudaFreeHost` calls are time-expensive ($\sim 100\mu s$), it is better to allocate arrays of the maximum size possible at the beginning and writing and reading continuously from them. Hence the arrays `length[]` and `offset[]` are needed, to keep track of the length and the position of an event in the big structure of arrays.

V. TESTS

Tests have been carried out on two machines.

Machine 1:

- CPU: Intel Xeon E5630
- GPU: NVIDIA Tesla Fermi C2050
- RAM: 12GB DDR2
- PCI Express v2

The NVIDIA Tesla C2050 card features clock speeds of 1.15 GHz core clock and 1.5 GHz on the 3GB of GDDR5 memory that runs on a 384-bit memory interface. Each SM, in turn, contains 32 CUDA processing cores to reach a total number of CUDA cores of 448.

Machine 2:

- CPU: Intel Ivy Bridge core i7 3770
- GPU: NVIDIA Kepler GTX 680
- RAM: 4x4GB Corsair XMS DDR3 2000MHz dual channel
- PCI Express v3 on Z77 chipset

Each SM is now a “next-generation Streaming Multiprocessor”, which Nvidia abbreviates as SMX; each SMX contains 192 CUDA cores, for a total of 1,536 cores in the entire Kepler GPU. The Kepler GTX 680 also contains four 64-bit memory controllers, operating at an overall data rate of 6,008MHz.

Machine 1 is running 64-bit Scientific Linux CERN 6.2, based on Red Hat Enterprise Linux 6.2, while Machine 2 is running 64-bit Fedora 17. This choice was obligatory due to the fact that support for NVIDIA Kepler GTX 680 and the

Fig. 3. Latency for a variable number of events inside each packet

Latency Stability

Fig. 4. Latency stability for a fixed number of events inside each packet

PCI Express v3 bus comes with kernel version superior to 3.3.

On both the machines NVIDIA CUDA Toolkit 4.2 and NVIDIA drivers v295.41 are installed.

In all the tests, the time spent copying the events from the host memory to the device memory is included, as well as the time spent copying the structure of the final results back to the host memory.

Since the photomultipliers divide the plane in discrete areas, all the calculations are made in integers and single precision floating point.

Figure 2 shows a typical GPU behaviour when processing a variable number of events. For a low number of events, the time spent is almost constant with respect to the package size since an increasing number of SMs is activated. Increasing the package size, two concurrent types of behaviour can be distinguished:

- an oscillatory one, due to the discrete nature of a GPU;
- a plateau, since when all the SMs are busy, the GPU is saturated.

The almost perfect factor $2\times$, both in throughput (Figure 2) and in latency (Figure 3), between the GTX680 and the

Tesla C2050 is due to the fact that this application is I/O bound, and the bandwidth of the PCI Express v3 bus is double the PCI Express v2 one. The tests show that on the Kepler GTX680 it's easier for the application to reach the aimed throughput (greater than 600MB/s) and latency (lower than 1ms) and the choice of the right packet size is not an issue.

Even if the measured latency for a packet is less than 0.5ms, a realistic stress latency test has been carried out. A realistic cycle of 15s ON and 15s OFF has been considered. 15s is the time interval between two beam spills. During the OFF part of the cycle, all the deallocations are called and the allocations, both in the GPU global memory and in pinned memory, are made again.

A histogram showing the latency times measured on the GTX680 during this test is shown in Figure 4. A tail at higher latencies exists but it is well within the desired limit of 1ms.

VI. CONCLUSION

We investigated some issues related to the implementation of a real-time GPU-based first-level trigger system. The latency and throughput of modern Kepler GPUs were found to be compatible with the requirements of a high-rate first-level trigger system such as the one of the NA62 experiment. For the RICH single ring fitting problem studied, considering a package containing 1000 events, a complex latency of 60μs and a throughput of 2.6GB/s were achieved. These results are well over the minimum requirements of a modern middle-sized high-rate HEP experiment such as NA62. More complex algorithms, required to deal with multiple rings and different trigger problems, can be more time consuming, but the approach looks quite promising. The next step is the implementation of a complete test system, involving multiple hardware Gb Ethernet data sources and links, on which continuous long-term performance can be studied. While we cannot expect the complete DAQ chain of HEP experiments to evolve into a fully commercial implementation (at least because of radiation issues in the earliest stages), a trend is firmly established towards standardized systems involving HEP-specific radiation-hard digitizing electronics and high-speed data links carrying the data to low-radiation areas where only commodity devices are used to perform all the required processing in real time. We believe the GPU-based approach described here to be a promising step for moving in such a direction.

ACKNOWLEDGMENT

The authors would like to thank...

REFERENCES

- [1] S. Jarp *et al.*, *Evaluation of Likelihood Functions for Data Analysis on Graphics Processing Units*, ipdpsw, pp.1349-1358, 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum, 2011.
- [2] W. Verkerke and D. Kirkby, *The RooFit Toolkit for data modeling*, proceedings of PHYSTAT05, Imperial College Press, 2006.
- [3] H. Perez-Ponce *et al.*, *GMC: a GPU implementation of a Monte Carlo dose calculation based on Geant4*, in 2011 IEEE Nuclear Science Symposium Conference Record (2011).

- [4] H. Perez-Ponce *et al.*, *Implementing Geant4 on GPU for medical applications*, in 2011 IEEE Nuclear Science Symposium Conference Record (2011).
- [5] S. Agostinelli *et al.*, *Geant4 - a simulation toolkit*. Nuclear Instruments and Methods in Physics Research A (2003).
- [6] G. Collazuol *et al.*, *Fast online triggering in high-energy physics experiments using GPUs*, Nucl. Instrum. Meth. A **662** (2012) 49
- [7] S. Gorbunov *et al.*, *ALICE HLT High Speed Tracking on GPU*, IEEE Transactions on Nuclear Science, Vol. 58, No. 4, August 2011.
- [8] *The NA62 collaboration - NA62 Technical Design*, NA62-10-07 CERN, Geneva (2010).
- [9] S. Adler *et al.* [The E949 Collaboration and E787 Collaboration], *Measurement of the $K^+ \rightarrow \pi^+ \nu \bar{\nu}$ Branching Ratio*, Phys. Rev. D **77** (2008) 052003 [arXiv:0709.1000 [hep-ex]].
- [10] F. Anghinolfi *et al.*, *NINO: An ultrafast low-power front-end amplifier discriminator for the time-of-flight detector in the ALICE experiment*, IEEE Trans. Nucl. Sci. **51** (2004) 1974.
- [11] B. Angelucci *et al.*, *Pion-muon separation with a RICH prototype for the NA62 experiment*, Nucl. Instrum. Meth. A **621** (2010) 205-211.
- [12] G. Haefeli *et al.*, *The LHCb DAQ interface board TELL1*, Nucl. Instrum. Meth. A **560** (2006) 494.
- [13] http://tdc.web.cern.ch/tdc/hptdc/docs/hptdc_manual.ver2.2.pdf M. Mota, J. Christiansen, JSSC, vol 34, no. 10, Oct 1999
- [14] J. F. Crawford, *A Noniterative Method For Fitting Circular Arcs To Measured Points*, Nucl. Instrum. Meth. **211** (1983) 223-225.