GPUs for fast triggering and pattern matching at the CERN experiment NA62

Gianluca Lamanna[a], Gianmaria Collazuol[b], Marco Sozzi[c]

[a]*Scuola Normale Superiore & INFN, Pisa, Italy*
[b]*INFN, Pisa, Italy*
[c]*University of Pisa & INFN, Pisa, Italy*

Abstract

In rare decays experiments an effective trigger is crucial to reduce both the quantity of data written on tape and the bandwidth requirements for the DAQ (Data Acquisition) system. A multilevel architecture is commonly used to achieve a higher reduction factor, exploiting dedicated custom hardware and flexible software in standard computers. In this paper we discuss the possibility to use commercial video card processors (GPU) to build a fast and effective trigger system, both at hardware and software level. The case of fast pattern matching in the RICH detector of the NA62 experiment at CERN aiming at measuring the Branching Ratio of the ultra rare decay $K^+ \to \pi^+ \nu\bar{\nu}$ is considered as use case although the versatility and the customizability of this approach easily allow exporting the concept to different contexts.

Key words: Trigger, Data Acquisition, Pattern recognition, GPU, Real-time

1. Introduction

In recent years the interest in using the GPU (Graphics Processing Unit, the processor used in standard video cards for PCs) for general purpose applications is growing due to the excellent performance shown in different fields of scientific computing. In particular, the parallel architecture of such processors, initially dedicated to image processing, is suitable to match the computing power requirements of many-body simulations or other problems in which massive computing is required. In the present generation a single GPU can provide up to 1 Teraflop computing power with a memory bandwidth in excess of 100 GB/s. This computing power is obtained thanks to the different resources allocation in the GPU with respect to the standard CPU: in the former more transistors are devoted to computing with respect to process control. Several projects in which a high computing capability is required employ clusters of PCs in which the computing load is shared between GPUs and CPUs. In particular in the field of high energy physics, several groups are pursuing the use of video cards for lattice QCD computations [1]. The use of GPUs presented here is different: we aim at employing the GPU for real-time decision in a trigger system for high energy physics experiments, studying both its use in a high-level software trigger and in a fixed-latency hardware trigger.

As a first application we have studied the possibility of fast pattern recognition in the RICH counter of the NA62 experiment using GPUs. In the following section the NA62 experiment and trigger will be briefly described, then GPUs and their application in the experiment will be introduced.

2. The NA62 experiment

The NA62 experiment at the CERN SPS aims at measuring O(100) $K^+ \to \pi^+ \nu\bar{\nu}$ events in two years of data taking, starting in 2012. The Branching Ratio (BR) for this decay mode is precisely predicted within the Standard Model (SM), with an irreducible error of few percent. This fact makes the $K^+ \to \pi^+ \nu\bar{\nu}$ unique both as a powerful test of the CKM paradigm and as a probe for new physics beyond the SM. Experimentally the detection of this process is very difficult due to the smallness of the signal (in the SM the expected BR is $(8.7 \pm 0.7) \times 10^{-11}$ [2]) and the presence of a large background, mainly from $K^+ \to \pi^+\pi^0$ and $K^+ \to \mu^+\nu$ (with BR of 20.7% and 63.5%, respectively). The present measurement of this decay channel is based on 7 candidates collected by E787+E949 Brookhaven experiments [3] leading to a value of $BR = (1.74^{+1.30}_{-0.89}) \times 10^{-10}$.

NA62 is a fixed target experiment in which a charged 75 GeV/c unseparated hadron beam, containing $\sim$ 6% of kaons, will be produced from 400 GeV/c protons from the SPS accelerator. Differently from previous experiments, NA62 will study the decay in flight, in a fiducial region $\sim$100 meters long, entirely in vacuum in order to reduce secondary interactions both from decay products and primary beam. In fig. 1 a schematic view of the experiment is shown. Background rejection and signal identification will be based on the high-resolution reconstruction of the decay kinematics: the four-momenta of both kaon and decay products will be measured with high resolution (Gigatracker and Straws). The non-kinematically-constrained background will be further identified using the information coming from the detectors used for particle identification (RICH, CEDAR and MUV) and veto (LAV, LKr, IRC and SAC). The NA62 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 good time res-

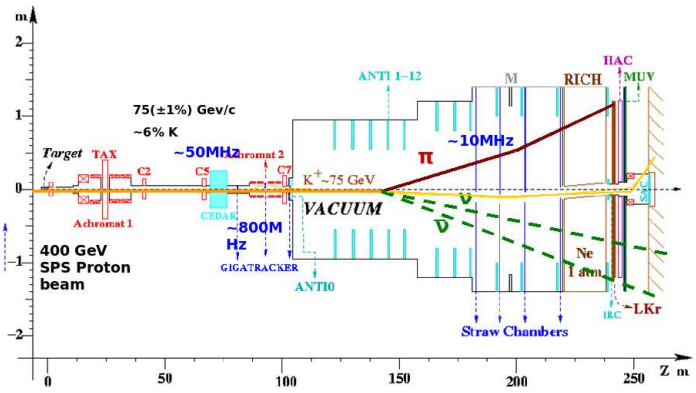

Figure 1: NA62 experiment layout.

olution. Čerenkov light is produced in an 18 m long, 3.7 m wide tube filled with neon at atmospheric pressure. The light is reflected by a composite mirror of 17 m focal length, and focalized on two separated spots at the opposite side of the detector. The two spots are equipped with photomultipliers (PM) (about 1000 PMs per spot) to reconstruct the rings. After amplification and discrimination in the NINO chip [4], the PM signal time is digitised by high resolution TDCs, as described in the next section. Given the quantum efficiency of the photomultipliers and the efficiency of light collection, a typical pion ring, for average accepted momentum, is identified with $\sim$ 20 firing PMs, as predicted by Monte Carlo and confirmed by test beams on a full-length prototype. The time resolution was measured to be better than 100 ps for all momenta in the considered range. Further information on the detector design and the experiment's physics goal can be found in [5].

2.1. The NA62 Trigger System

The goal of collecting 100 SM events in a reasonable amount of running time can be achieved using a very intense beam and a reliable data acquisition (DAQ) and trigger system. 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 lossless 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 NA62 trigger and DAQ system are integrated in a unified completely digital system (TDAQ): the readout data, stored in large buffers waiting for trigger decisions, is exactly the same as used to construct the trigger primitives. It should be noted that a completely "triggerless" approach in which the data are entirely read out in PCs, avoiding any hardware level data reduction, cannot be afforded for an experiment with thousands of channels and a rate of tens of MHz. 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 will be software-based: the L1 decision is based on single-subdetector reconstructed quantities, while the L2 decision is taken on the fully reconstructed event with high resolution. The L0 is based on the TELL1 [6] (developed by EPFL for the LHCb experiment), a general purpose acquisition board in which 5 FPGAs allow a fully customisable configuration. Thanks to a RAM of 384 MB, the TELL1 can store data in a first buffer stage, waiting for the trigger decision delivered to the board through the CERN standard TTC [7] interface. The TELL1 board is going to be redesigned in order to use more performant FPGAs and larger, faster buffers. In the TELL1 up to four daughter boards can be mounted. For the majority of the detectors in NA62, time is the most important information to be provided. Therefore a daughter board with 4 HPTDC ASICS [8] has been designed, in order to have 512 TDC channels per TELL1 with 100 ps time resolution. The TELL1 allows to build in FPGAs the trigger primitives combined in the L0 trigger processor (L0TP). The L0 trigger decision is based on the presence of a charged particle in the RICH and veto conditions on LKr, LAV and MUV, to obtain a reduction factor of 10. The trigger decision is broadcasted synchronously to all TELL1 acquisition boards with a latency of 1 ms. The events of interest will go into the L1 via ethernet connection. Due to large amount of data to be processed in a reasonable time, the number of PC cores at the L1 will be quite large. After this level of selection the data from all the detectors will arrive at L2 through a network switch; the event will be fully reconstructed in order to apply a tighter selection based on the full kinematics. 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).

3. The GPU (Graphics Processing Unit)

In recent years efforts have been devoted to the development of powerful processors dedicated to graphical application. The peculiarities of problems related to 3D, rendering and, in general, image processing, drove the development of parallel architectures SIMD (Single Instruction Multiple Data) type. In such an architecture the same algorithm is applied simultaneously to different sets of data, using several computing cores running at the same time. Different levels of parallelization could be implemented in such a scheme by grouping the multicore structure in different layers. Thanks to this structure the GPU's computing power can easily exceed the Teraflop level and improvements are related to the total number of transistors used, rather than the clock speed (as in the CPU case). Recently interest arose for the use of GPU in general purpose applications (GPGPU) [9] outside the field of imaging processing, with particular focus on high performance computing for scientific problems. Several examples of applications can be found in literature going from seismology to medical physics and calculations in lattice QCD [1]. While GPGPU applications related to high-performance computing tasks are nowadays becoming commonplace, so far applications in which GPUs are used for hard real-time tasks have been hampered by the large intrinsic latencies related to the relaxed requirements of the graphic applications on this point. However, the continued progress in speed is such that GPU latencies are becoming compatible also with such real-time tasks. The two main vendors (ATI and NVIDIA) are involved in the task to provide video cards

in which the GPU can work as a co-processor for vectorizable problems. In particular NVIDIA [10] proposes a comprehensive and consistent approach to general-purpose scientific computation. The NVIDIA Tesla C1060 card employed in the work described in this paper houses one GT200 GPU with 240 computing cores and 4 GB DDR3 memory with a bandwidth of ~102 GB/s.

4. A trigger system based on GPUs

In a standard trigger system for a high energy physics experiment, the complexity in primitives construction and trigger decision is limited by the time available as defined by latency requirements. Usually in trigger levels with fixed small latency, the trigger primitives are quantities related to multiplicity and hit patterns. The trigger decision is defined with rough conditions, not allowing high rejection factors and selection power. The use of computing units (possibly housed in a standard PC) in the definition of high quality trigger primitives and trigger decisions could be exploited to build more selective trigger systems. In this regard some critical points must be taken into account:

- data transfer from readout to PC: fast, reliable and time-determistic dedicated links must be employed in order to exploit a large bandwidth;
- small latency and high rate: high computing power to have compact and cheap systems, avoiding large and expensive switched networks;
- stable latency: small spread in the execution time without non-Gaussian tails.

The last two points, in particular, could be addressed in a GPU based system: GPUs, indeed, have both a lot of computing power and a quasi-deterministic behavior. In contrast, to date the use of standard CPU mother boards to manage commercial GPUs cannot be avoided. The behavior of standard CPUs under the control of a standard operating system is far from being deterministic: precautions to limit any time variability due to shared CPU usage must be implemented, such as the use of real-time operating systems, “smart” network cards to manage the network protocol and DMA (Direct Memory Access) mechanisms, in order to avoid time losses in copying data to the PC memory.

5. GPU used for fast pattern matching in the NA62 RICH detector

In many cases the definition of trigger primitives can be reduced to pattern recognition issues. This is the case for charged-particle track identification in magnetic spectrometers, trajectories in silicon strip trackers or photon rings in Čerenkov detectors. The RICH counter in the NA62 experiment falls into this last category: the center and the radius of the Čerenkov rings in the detector are related to the angle of the particle and its velocity, respectively. This information could be employed at the trigger level to increase the purity and the rejection power of many triggers of interest. The power and the speed of GPUs can be an interesting solution to this problem.

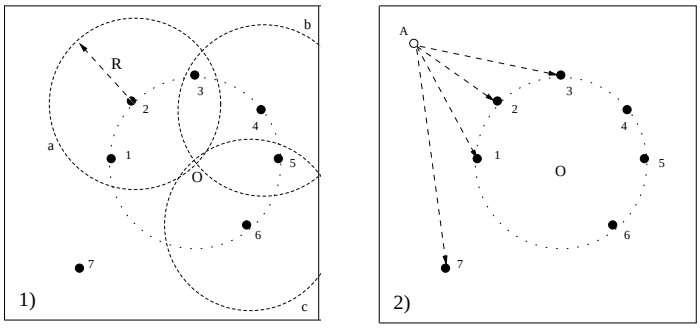

Figure 2: Ring-finding algorithms implemented in GPU. Black dots are hit PMs. 1) In the GHT algorithm probe circles (a, b and c) for a fixed radius are exploited to find the Čherenkov ring center. 2) In the POMH and DOMH algorithms, for each potential centre A the distances to all hits are computed to fill a histogram.

In order to test the feasibility and the performance, as a starting point we have implemented three algorithms for ring finding in a sparse matrix of 1000 points (for convenience placed at the center positions of the PMs) with 20 firing PMs (“hits”) on average. To better reproduce the conditions of interest we used the Geant4-based [11] official NA62 experiment Monte Carlo simulation.

The specific GPU architecture must be carefully considered in order to achieve the best performance. The computing cores are grouped into “multiprocessors” sharing memory and instruction pool. The access to on-chip memory is very fast (up to 1 TB/s) if read and write conflicts are carefully avoided. The processes (the “threads”) running in each core must be synchronized at the multiprocessor level to maintain concurrent execution without divergences and consequent partial serialization.

The first algorithm we tested is based on a Generalised Hough Transform (GHT) (fig. 2). In this approach each hit is considered as the center of a probe circle with fixed radius. The point with the largest number of intersecting circles, varying the radius over the range 5 to 11 cm in steps of 2 cm, is considered as the center of the Čherenkov ring. The main limitation for this algorithm is due to the amount of on-chip fast memory available. This limits the size of the tridimensional space (center position and radius) used for the maximization procedure. The advantage is that for each event a small number of threads (proportional to the number of hits) have to run concurrently on the GPU.

In the other two algorithms each point in a fixed grid is considered as candidate for a center (fig. 2). A histogram of the distances between the point and all the hits is constructed in order to identify the true center and the corresponding radius. The difference between the two algorithms, called Problem-Optimized Multi Histograms (POMH) approach and Device-Optimized Multi Histograms (DOMH) approach, respectively, is in the management of the parallelization structure in the GPU. In the POMH case each core has to make very simple operations (a distance calculation in a plane) but the whole proces-

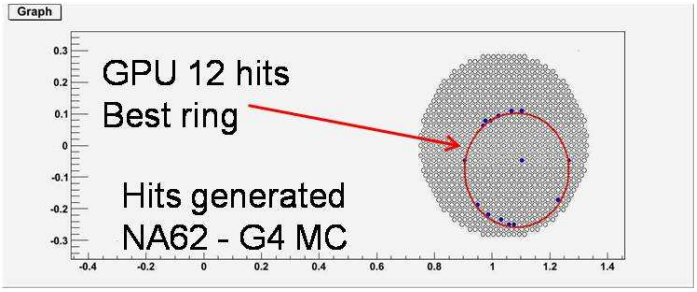

Figure 3: Example of identified Čherenkov ring obtained using a GPU.

sor is employed for the same event; in DOMH the single core has to perform a trickier task in order to optimize the number of concurrent processes with the read and write procedure in the fast shared memory and to allow processing of multiple events simultaneously in the same chip. In this second case resources are used more efficiently: both the possibility of parallelizing the algorithm and of processing several events are exploited.

DOMH algorithms show the best performance and allow to find a ring in 3.1 μs (with negligible non-Gaussian tails) when a packet of 1000 events is sent to the GPU for processing (fig. 3). Thanks to the high bandwidth provided by the PCI-E bus (4 GB/s), the data transfer from an acquisition card (such a Gigabit ethernet card) to the GPU through the internal PC bus, should not be an issue for the case of NA62. The averaged event size for the RICH is ~ 200B; assuming a maximum rate of 10 MHz the bandwidth is ~ 2 GB/s in input at L0 and a factor of 10 less at L1. In addition, data transfer and computation in the GPU are managed concurrently: no time is lost either in data preparation for processing or for transferring back the results from the video card to the PC.

6. Integration in the NA62 trigger system

The possibility to perform fast and powerful computations at the software level is a key point for reducing the size of the on-line PC farm and the processing time. On the other hand, the possibility to define more complex quantities (like invariant masses, momenta, etc.) at the earliest time (L0) could be very useful for achieving a more efficient on-line selection and a more effective rejection. In the NA62 trigger architecture GPUs can be easily integrated in the L1 and L2 software levels based on PCs. The maximum rate for the input at L1 is less than 1 MHz in the standard L0 configuration and the latency is not an issue. The software levels could easily take advantage of the use of graphics processors in the computing stage. In case of L0 the advantage in using GPUs would be twofold: trigger primitives could be computed with high resolution (comparable to off-line) and triggers with higher purity and efficiency could be defined at the lowest levels, allowing to use the readout bandwidth with additional physics and control triggers of interest. Use of GPUs in the L0 is very challenging because of the high input rate and the required small and stable latency. For instance, the time for ring identification, as in the example discussed above, should be 100 ns, as the rate at L0 is 10 MHz. The best preliminary result shown here (3.1 μs) does not yet allow to implement the L0 trigger of NA62 on a small number of GPUs but the following considerations should be taken into account:

- there is still room for improvement in the presented algorithms, and new algorithms are under test;
- the system employed is based on a single GPU. The total load can be easily distributed over several GPUs.
- the next generation of graphics processors (now on the market) provides better performance (by at least a factor of 2).

Given these considerations, the possibility of using GPUs directly at the hardware level of the trigger is still very attractive and is being actively proved.

7. Conclusions

The use of commercial video cards in a high energy physics trigger system is a very interesting possibility for several reasons. The power of GPUs exceeds that of CPUs by orders of magnitude, which allows to have very versatile and compact computing units to address problems of on-line event selection. In particular the use of GPUs is now becoming possible for real-time applications thanks to the deterministic behavior of the computation and the increase in computation speed. Thanks to the use of components designed for commercial sectors with large markets, this solution appears to be very cheap compared to other ones based on custom hardware. In addition, a system based on GPUs benefits directly from the continuous technological progress required by video games and image processing applications.

The use of GPUs at both the “hardware” and “software” levels allows to define a new architecture for trigger systems where the hardware (custom) part is reduced to digitization and buffering while the whole logic, based on digital data, is performed in software. Such a scheme can be easily adapted to any high-energy physics experiment to increase the on-line selection power and to decrease the total cost.

References

[1] G. I. Egri, Z. Fodor, C. Hoelbling, S. D. Katz, D. Nogradi and K. K. Szabo, Comput. Phys. Commun. **177** (2007) 631 [arXiv:hep-lat/0611022].
[2] J. Brod and M. Gorbahn, Phys. Rev. D **78** (2008) 034006 [arXiv:0805.4119 [hep-ph]].
[3] S. Adler *et al.* [The E949 Collaboration and E787 Collaboration], Phys. Rev. D **77** (2008) 052003 [arXiv:0709.1000 [hep-ex]].
[4] F. Anghinolfi, P. Jarron, F. Krummenacher, E. Usenko and M. C. S. Williams, IEEE Trans. Nucl. Sci. **51** (2004) 1974.
[5] NA62 Collaboration, CERN-SPSC-2007-035 (proposal).
[6] G. Haefeli, A. Bay, A. Gong, H. Gong, M. Muecke, N. Neufeld and O. Schneider, Nucl. Instrum. Meth. A **560** (2006) 494.
[7] B. G. Taylor, IEEE Trans. Nuclear Science, Vol. 45, No. 3, (June 1998)
[8] M. Mota, J. Christiansen, JSSC, vol 34, no. 10, Oct 1999
[9] `http://www.gpgpu.org/`
[10] `http://www.nvidia.com/cuda/`
[11] S. Agostinelli *et al.* [GEANT4 Collaboration], Nucl. Instrum. Meth. A **506** (2003) 250.